सत्यमेव जयते

INDIAN AGRICULTURAL
RESEARCH INSTITUTE, NEW DELHI

I.A.R.I.6.
GIP NLK—H-3 I.A.R.I.—10-5-55—15,000

PROCEEDINGS

OF THE

ROYAL SOCIETY OF LONDON

SERIES A. MATHEMATICAL AND PHYSICAL SCIENCES

VOL 187

LONDON
Printed and published for the Royal Society
By the Cambridge University Press
Bentley House, N.W.1

13 December 1946

Printed in Great Britain at the University Press, Cambridge
(Brooke Crutchley, University Printer)
and published by the Cambridge University Press
Cambridge, and Bentley House, London

Agents for Canada and India: Macmillan

CONTENTS

SERIES A VOL 187

No. A1008. 8 October 1946

PAGE

Motion of waves in shallow water. Interaction between waves and sand bottoms. By R. A. Bagnold, F.R.S. (Plates 1–4) 1

(With a note on R. A. Bagnold's empirical formula for the critical water motion corresponding with the first disturbance of grains on a flat surface. By Sir Geoffrey Taylor, F.R.S.) 16

The photochemical polymerization of methyl vinyl ketone vapour. By T. T. Jones and H. W. Melville, F.R.S. 19

The free radical polymerization of the vapours of certain vinyl derivatives. By T. T. Jones and H. W. Melville, F.R.S. 37

The adsorption of vapours on mercury. I. Non-polar substances. By C. Kemball and E. K. Rideal, F.R.S. 53

The adsorption of vapours on mercury. II. The entropy and heat of adsorption of non-polar substances. By C. Kemball 73

Two-dimensional boundary-value problems of elasticity. By Bibhutibhusan Sen 87

The mechanism of ionospheric ionization. By R. v. d. R. Woolley . . . 102

Stability of viscous flow between rotating cylinders. I. By D. Meksyn . . 115

SERIES A VOL 187

No. A1009. 22 October 1946

The kinetics of the reactions of the steam-carbon system. By J. Gadsby, C. N. Hinshelwood, F.R.S. and K. W. Sykes 129

On lattice points in n-dimensional star bodies. I. Existence theorems. By K. Mahler 151

The Raman spectrum of rock-salt. By R. S. Krishnan. (Plate 5) 188

On Dirac's theory of quantum electrodynamics: the interaction of an electron and a radiation field. By C. Jayaratnam Eliezer 197

The application of quantum electrodynamics to multiple processes. By C. Jayaratnam Eliezer 210

Krypton vacuum wave-length measurements. By T. A. Littlefield 220

The distribution of stress in the neighbourhood of a crack in an elastic solid. By I. N. Sneddon 229

No. A1010. 5 November 1946

PAGE

The basic reactions in the upper atmosphere. I. By D. R. Bates and H. S. W. Massey, F.R.S. 261

The effect of an electric field on the viscosity of liquids. By E. N. da C. Andrade, F.R.S. and C. Dodd 296

The thermochemistry of carbon: valence states, heats of sublimation and energies of linkage. By L. H. Long and R. G. W. Norrish, F.R.S. 337

The experimental study of the blast from bombs and bare charges. By G. Grime and H. Sheard 357

Elastic constants of diamond. By S. Bhagavantam and J. Bhimasenachar . . 381

Algebra related to elementary particles of spin $\frac{3}{2}$. By B. S. Madhavarao, V. R. Thiruvenkatachar and K. Venkatachaliengar 385

Some thermodynamical properties of gaseous sulphur dichloride. By C. A. McDowell and E. A. Moelwyn-Hughes 398

No. A1011. 13 December 1946

The mechanism of ionospheric ionization. II. By R. v. d. R. Woolley . . . 403

The statistical aspect of fatigue of materials. By A. M. Freudenthal . . . 416

Kinetic friction in or near the boundary region. I. Apparatus and experimental methods. By B. Chalmers, P. G. Forrester and E. F. Phelps. (Plate 6) . 430

Kinetic friction in or near the boundary region. II. The influence of sliding velocity and other variables on kinetic friction in or near the boundary region. By P. G. Forrester 439

A cloud-chamber investigation of penetrating showers. By G. D. Rochester. (Plates 7–9) 464

Stability of viscous flow between rotating cylinders. II. Cylinders rotating in opposite directions. By D. Meksyn 480

Stability of viscous flow between rotating cylinders. III. Integration of a sixth order linear equation. By D. Meksyn 492

Index 505

Motion of waves in shallow water
Interaction between waves and sand bottoms

By R. A. Bagnold, F.R.S.

(*Received* 14 *August* 1945—*Read* 31 *January* 1946)

With an additional note by Sir Geoffrey Taylor, F.R.S.

[Plates 1–4]

The loss of energy of a travelling water wave, due to the mechanism of the formation of sand ripples and water vortices on a sandy bed, becomes of practical importance when models are used to predict full-scale foreshore movements.

On the assumption that the bottom-water oscillation is nearly simply harmonic, the mechanism was studied by oscillating a section of bed through still water.

The pitch, p, of the sand ripple formed was found to vary as the square root of the grain diameter, independently of the speed and of the grain density, for amplitudes, R, of water motion exceeding this pitch. But for smaller amplitudes the pitch shortens with decreasing amplitude of movement.

The mean drag coefficient, k, in the case of artificial rigid ripples, was measured directly. For R/p less than unity, k remains constant. For R/p greater than unity, k was found to vary as $(R/p)^{-0\cdot75}$. These results are compared with the case of steady flow.

The critical water speeds and amplitudes for first disturbance of grains on a smoothed surface was also measured, over a wide range of grain diameters and densities. The results conform closely to a simple empirical expression.

1. INTRODUCTION

The success of many major maritime engineering projects depends on the correct prediction of the movement of the bottom under the action of waves. Were the laws known which govern the interaction between the bed material and the oscillating motion of the water in contact with it, the use of models would be of far greater help than is at present the case. For though model waves over a smooth rigid bed run remarkably true to scale, serious discrepancies occur when one tries to imitate their effect on loose granular bed material. The quantity of sand transported is too little; the area where the bed is active is too restricted; and the bed is corrugated with ripples of a size out of all proportion to their full-scale prototypes. These comparatively big ripples make the drag between the bed and the water too great, and the waves become exhausted too soon.

The object of the experiments to be described was to obtain quantitative data (*a*) on the size and character of the sand ripples made by waves, (*b*) on the drag to which these ripples give rise, and (*c*) on the minimum oscillating water motion required to disturb the sand in the first instance.

The experiments were confined to the simplest case of a continuous train of symmetrical waves travelling over a horizontal bed. They are intended as a first instalment of a research into the movement of sand by water motions in which wave action predominates.

2. Bottom-water motion

Classical wave theory, though it neglects drag, was assumed to give with sufficient accuracy the water motion near the bottom. For waves in deep water and also for those shallow-water waves in which the wave height is small compared with the depth, the horizontal oscillations of the water near the bottom are known to approximate closely to simple harmonic motion. For shallow-water waves of extreme height a connected train of solitary waves of greatest height may be taken, as worked out by McCowan (1891). Here, too, when a suitable steady reverse motion is superimposed on the whole body of the water in order to approximate to the practical case of a channel of finite length, the remaining motion at the bottom deviates surprisingly little from simple harmonic. It is only when the repetition distance exceeds the 'wave-length' $2\pi/m$ that large deviations occur. The horizontal acceleration at the bottom appears never to exceed $g/6$ provided the wave is steady.

Hence simple harmonic bottom-water motion of amplitudes less than $g/6\omega^2$ (where the angular velocity $\omega = 2\pi/\text{wave period}$) appears to cover a wide range of practical conditions.

3. Nature of the experiments to be described

In view of the above it was felt that the results of observations, made far more conveniently by oscillating the bottom harmonically through still water, would not differ appreciably from the real effects on a still bottom produced by water motion provided the accelerating form on the grains via the bed, due to the driving mechanism, is small compared with the fluid force acting on them. All the results given and discussed in this paper have been obtained in still water.

The two sets of experiments were carried out between October 1944 and February 1945 in the hydraulics laboratory at Imperial College with the help of Dr C. M. White.

In experiment 1, in which was observed the behaviour of actual sands, the sand was contained in a cradle suspended from a pivot in the roof so that it could be oscillated through a circular arc in still water in a narrow tank. (Incidentally, this tank happened by chance to be one of those used years ago by Mrs Ayrton for her qualitative study of water-sand ripples.) The oscillation was maintained by means of a motor-driven crank, as shown in figure 5*a*, plate 1.

For each of a range of different semi-amplitudes R, from 25 to 0·5 cm., a series of runs were made at different angular speeds ω, increasing from the lowest speed at which each sand began to move over the bed, up to the point at which the acceleration of the cradle at the end of the stroke reached such a value $\mu g\frac{(\sigma-\rho)}{\sigma}$ that the whole sand mass started sliding on the steel cradle-floor. Between these limits a wide range of ripple phenomena was observed.

In experiment 2 (figure 5*b*, plate 1), by which the ripple drag was measured, a celluloid plate, of submerged dimensions 1 m. long and 50 cm. deep, to which fixed imitation ripples were attached, was hung vertically in a large tank of water from a horizontal runway. The plate was oscillated by the same driving mechanism, from which, however, the motor was disconnected and replaced by a winding drum and a wire to which weights could be hung.

This apparatus, motor driven, was also used to take still and cine-pictures of the water motion near the plate, as indicated on the free surface by a scattering of aluminium powder. The cameras were mounted on the plate carriage and moved with it.

4. GENERAL OBSERVATIONS ON THE WATER MOTION OVER THE RIPPLE

In figure 6, plate 2, where R was 20 cm., the ripple pitch p, 10 cm. and ripple height h (trough to crest), 1·5 cm., the phase of the stroke in which each exposure was made is indicated by the pointer, end of stroke being at the bottom. The series shows one stroke (half a completed oscillation) during which the pointer rotated anticlockwise through a complete revolution, the ripple (and camera) moved to the left and the main water-mass to the right.

In interpreting the series it should be remembered that the camera moved with the ripples, and though the small eddies within the ripple trough and travelling with it appear sharply in each photograph, the large free vortices are only visible clearly as such when they happen to be at rest relatively to the camera, i.e. at the end of the stroke. They are, however, there all the time, and their ghosts can be distinguished by the waviness of the particle streaks.

The water motion differs according as the stroke length $2R$ is greater or less than the ripple pitch p (crest to crest). If it is greater (high mid-stroke velocities and low end-stroke accelerations), the trough is occupied during most of the stroke by small eddies which have developed from the lee crest, and we seem to have the beginnings of successive true, though incipient, stead-flow boundary layers. For strokes many ripples in length the little eddies may grow, break up and drift away and be replaced by others several times before the stroke ends.

But as the stroke is shortened, the formation of the eddies inside the troughs is delayed till a later and later phase, and when $2R$ is less than p the troughs are empty of eddies till the last quarter is reached.

During retardation at the end of the last quarter of the stroke, a single eddy in the middle of the trough rapidly grows into a very large vortex, which fills the whole trough. At the end of the stroke, or in the case of very short strokes at the beginning of the next return stroke, its inertia causes the vortex to lag behind, so that it leaves the middle of its trough and mounts up the slope past the crest and out into free water. During successive subsequent strokes it is forced farther and farther outwards away from the surface as new vortices are formed and ejected after it. In the first photograph the vortex appearing centrally over the ripple crest is an old one ejected at the end of the previous stroke. The remains of older ones can also be seen farther out. The general outward drift of the vortices produces a slow cellular circulation in the mass of the water, which extends to a great distance from the ripple surface; and if the terminal velocity of fall of the sand grains is slow, the vortices carry up curtains of suspended grains with them.

Time has not allowed of a detailed study of the effect on the water motion of varying the shape of the ripple section. In the artificial fixed ripples used, the ratio of wave-length to height was 6·7. It will be apparent from what follows that the big end-of-stroke ripple vortex will not form if this ratio is much increased. The formation of the ripple vortex seems to be confined to oscillatory motion, and there is no evidence for the existence of vortices of such a size under conditions of steady flow. The frictional drag between plate and water is discussed in § 9.

5. RIPPLE FORMATION UNDER OSCILLATORY WATER MOTION

The rolling-grain ripple. The relations between the sand-grain size and density, the repetition distance or pitch p of the ripple, and the elements ω and R of water motion producing it was complicated by the existence of at least two distinct types of ripple, each having its own formatory mechanism. Though the two ripple types may in some cases be similar in appearance, they differ greatly in character. The similarity in appearance between the two kinds of ripple, and between these and the profile of high-crested water waves is due respectively to the constancy of the angle of repose to which sand grains are tipped over the ripple crest in the two cases, and to the fact that this angle of a little over 30° makes at the crest, quite fortuitously, the precise angle of 120° which is the limiting angle of the water wave of greatest amplitude.

At the critical speed of the water motion at which grains on a smoothed surface first begin to move (see § 11), grains start to be rolled to and fro over the surface, but are not lifted off it. Since the grain movement is here confined to that phase of the stroke at which the relative water-sand velocity at the surface is greatest (which occurs at about quarter-stroke), the length of the grain path remains short. It lengthens, however, as the oscillation speed is raised.

Though initially distributed at random, the rolling grains become organized as time goes on, and tend to come to rest in parallel transverse zones. More grains reach these zones than leave them, so there is a progressive congregation of grains in them, and the zones soon become little wavy ridges a few grains high, whose crests sway from lee side to lee side during successive stroke reversals. Each lee side becomes a miniature scree at the angle of repose. As the ridge grows it shelters from the water action a wider and wider strip of flat surface on its successive lee sides; and when the sheltered area extends as far as the next ridge no further grain movement can take place anywhere but on the ridge itself. Hence, since the ridges can now collect no more grains, they cease to grow, and the arrangement becomes stable. The repetition distance is evidently the width of the flow shadow of the ridge, and depends on its height.

If the speed is raised, the increased water speed over the intervening flat strips sets more grains rolling, and these in turn become trapped whenever they reach a ridge. So the ridges grow again until a new state of stability is reached. Adjustment of the repetition distance between the ridges takes place by redundant ridges joining up and amalgamating.

The rolling-grain ripple occurs on all sands, but the profile varies. With fine grains the intervening surface remains flat. But with larger grains it becomes a nearly circular arc of large radius. The essential distinguishing feature of this type of ripple is, however, the entire absence of grain movement within the troughs, whose surfaces remain undisturbed either by water flow or by the impact to descending grains—for no crest grains leave the surface. The length-height ratio is evidently too big for the formation of the ripple vortex.

The rolling-grain ripple appears to be stable, for a given oscillation amplitude, between the critical speed of first movement on a smoothed surface and about double that speed, and within this range the vortex ripple will not form at all, provided there are no irregularities anywhere on the surface higher than say 20 grain diameters.

But if the speed exceeds twice the critical speed, the steep lee slopes reach such a height that an abrupt change takes place in the character of the water motion; the ripple vortex suddenly appears, and the whole regime breaks down.

The vortex ripple. The breakdown begins at one spot, where the crest happens first to exceed the critical height. From this spot the new regime of vortex ripples spreads rapidly over the rest of the surface like a disease. Successive stages are shown in figure 7, plate 3. Here the breakdown was started artificially by placing a small heap of sand, about 30 grain diameters high, on the surface. But any kind of surface feature will suffice—a pebble dropped on it, for instance, or a steep-sided dent made in it.

Indeed, it is not even necessary that the oscillation speed should have attained that needed for the spontaneous breakdown of the rolling-grain ripple. For the vortex-ripple mechanism will operate, and the ripples develop, from any sufficiently large surface feature at speeds even below that needed to start grain movement on a smoothed-out surface.

The operation of the vortex-ripple mechanism appears to be as follows. At the end of the stroke when the fully grown vortex begins to overrun the ripple, it comes into contact with the sand at the ripple's foot. Grains are scooped out from here and shot upwards parallel with the surface of the slope, as if the sand at the bottom had been touched by a spinning wheel. The path of the bulk of the lifted grains is such that they come to rest just at the crest of the ripple; and it seems to be this fact that maintains the ripple system in equilibrium and controls the size of the ripple.

With grains of a high terminal velocity of fall, none of them overshoot the crest into the water above, and there is therefore no suspended sand cloud. But with fine or with light grains a proportion of them get involved in the water of the vortex, and are carried up with it.

After the grains from the bottom have been lifted to the crest there is, except with very short strokes, a short pause till the vortex on its upward journey parallel with the ripple slope has passed the crest and become separated from the ripple altogether. By this time the ripple has just begun to move backward in the reverse direction. A jet of water now begins to squirt down the outward slope of the ripple, between the crest and the departing vortex. This jet flicks a volley of grains from the crest downwards parallel with the outward ripple slope. The grains impinge violently on the sand surface at the bottom, ricochet off it, scoring out a hollow where they strike, and finally come to rest beyond, where they gradually collect and build up the ridge of a second ripple.

As the return stroke proceeds no further grain movement takes place in the trough until the end of the stroke, when new vortices perform the same operation in the opposite direction. But during the stroke, while there is no large vortex in the ripple trough, the velocity of the water flow over the crest tends to flatten it by dislodging the grains at the sharp cusp and rolling them over down the lee side.

This velocity effect at mid-stroke becomes more pronounced as the stroke is made longer, so that the ripple height decreases. When the stroke length is very long and the velocity high the crests become so flat and rounded that the limiting wave-length/height ratio is reached at which the restoring end-of-stroke vortex cannot develop.

With very short strokes, on the other hand, there is no horizontal flow over the crest at any part of the stroke, and the crests remain sharply cusped, as in plate 3.

6. THE PITCH OF VORTEX RIPPLES

The first attempts to measure the pitch of the ripples gave inconsistent results because it was not realized that the ripples are capable of considerable compression when overcrowded. When generated directly from a smooth sand surface at speeds exceeding the breakdown value for the rolling-grain ripple (i.e. exceeding twice the critical speed), vortex-ripple nuclei are formed at several haphazard places at once. As each of these develop and multiply, the separate groups compete with one another for space; and the resulting pattern, though regular and apparently stable, has in reality a shorter pitch than would be the case if each ripple had adequate 'living room'. Thinning out can only occur if two ripple crests happen to incline towards one another and to coalesce. So, since the transverse arrangement seems to be very stable, accidental coalition can occur but rarely and one has to wait a very long time.

This elasticity is no doubt possible owing to the fact that each ripple ridge together with its alternate lee-side vortices comprises a complete unit capable of separate existence. At the true natural spacing the mean points at which the vortices scoop out grains from the bottoms and throw them up to the crests coincide with the centres of the troughs. Overcrowding merely results in these points being pushed off the trough centres towards the feet of the opposite ridges.

This uncertainty as to the true natural pitch was finally overcome by ensuring that the ripple system started from a single nucleus only. This was done by working at very low speeds—the critical speed for initial smooth-surface grain movement was taken as a standard—and by starting the ripple mechanism from a single artificially made transverse ridge. From this nucleus the ripples can multiply freely in both directions. The final natural pitch was assumed to have been attained when three adjacent ripples had grown to equal size. Ripples formed at higher speeds were started by the same method.

In all cases it was found that the pitch was independent of the speed of the oscillation. The experimental pitch measurements for sands of various grain diameters and densities are plotted in figure 1 against the oscillation amplitude R. A wide range of grain diameter and of grain density was used. The diameter varied between 0·25 and 0·009 cm., and the densities used were those of steel ($\sigma = 7{\cdot}9$), quartz ($\sigma = 2{\cdot}65$) and coal ($\sigma = 1{\cdot}3$).

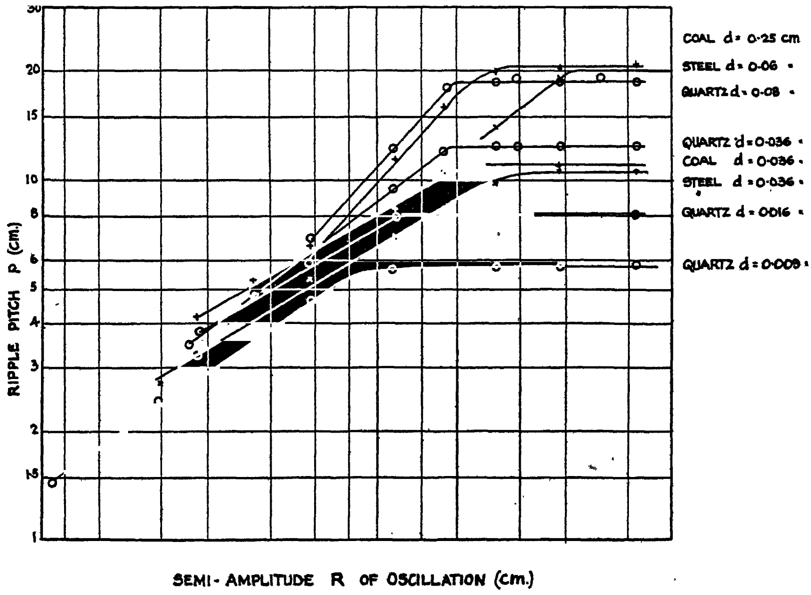


FIGURE 1

It will be seen that as the amplitude R of the stroke was decreased from the experimental maximum of 32 cm. the pitch p of the ripple, measured from crest to crest, remained constant at a value P which seems to depend only on the characteristics of the sand. But with the exception of the two biggest grain sizes, coal 0·25 cm. and quartz 0·08 cm., the pitch length began to fall suddenly as soon as the ratio R/P had reached unity, and for smaller values of R this ratio remained approximately constant.

From table 1 (in which values of P and d are given in cm.) it appears that the natural pitch P of the sand is independent of the grain density and varies nearly as the square root of the diameter. The range of values of ω and R for which the experimental values of P were found to hold extended from the lowest value of ω for initial grain disturbance, as given in § 11 up to the limits discussed in § 8.

TABLE 1

steel			quartz			coal		
d	P	$P/\sqrt{d}$	d	P	$P/\sqrt{d}$	d	P	$P/\sqrt{d}$
0·06	20·0	82	0·08	18·7	66	0·25	20·5	41
0·036	10·5	55	0·036	12·3	64·5	0·036	11·0	58
			0·016	8·0	63·5			
			0·009	5·8	61			

The above measurements were made on the pitch of ripples in sands of nearly uniform grain size. In order to test the effect on the pitch of non-uniformity of size, a quartz sand was made up in which the grades fell off very gradually on both the smaller and larger sides of the dominant diameter of 0·036 cm. It was found that although the size distribution was considerably wider than most sands found in nature, the ripple pitch still followed exactly the same curve as that for the uniform sand of 0·036 cm. diameter shown in figure 1.

The absence of any appreciable variation with the grain density is curious. Of the two most likely ways in which the diameter alone could become a controlling factor, surface roughness within the ripple trough and permeability of the material, there is a little evidence in support of the former. For in the case of the sand of mixed grain size which had the same value of P as that of the uniform sand of the dominant diameter, there was a distinct segregation of the grains; the hollow of the ripple was covered with those of the dominant size, whereas the smallest and largest grains collected together to form the crests.

The pitch/height ratio of vortex ripples reaches a minimum value of between 4·5 and 5 near the bend of the curves in figure 1. It increases as the amplitude is lengthened, till at the experimental limit of $R = 32$ cm. it had risen to nearly 8. The ratio appears to remain constant for shorter amplitudes, but at very short amplitudes measurement was complicated by the appearance in the ripple troughs of the features described in the next section.

7. THE 'BRICK PATTERN' AT VERY SHORT STROKE AMPLITUDES

A curious modification of the vortex ripple appeared with all sands, independently of the working speed, when R was reduced to about $P/6$. Longitudinal and equidistant bridges were formed spanning the ripple troughs from crest to crest. The transverse spacing between them was very regular, being either one, two or three ripple pitches; and the positions of all the bridges across one ripple trough always 'broke joint' evenly with those across the troughs on either side, so that a perfect diagonal pattern was created. The effect is shown in figure 8, plate 4. Sometimes the bridges rose higher above the surface than the ripple crests themselves. The water

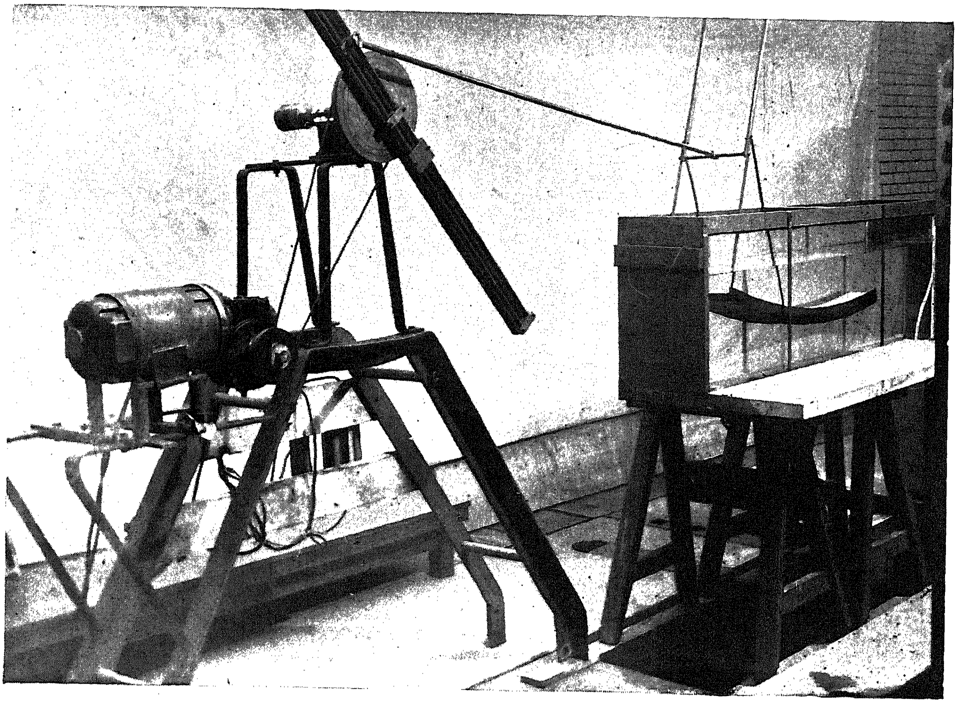


FIGURE 5 (*a*)

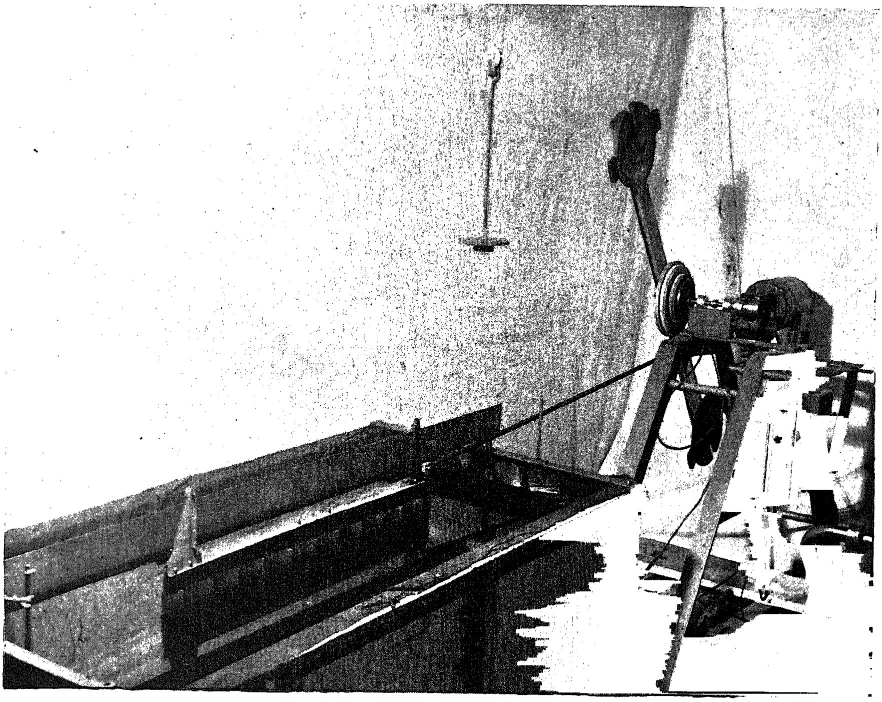
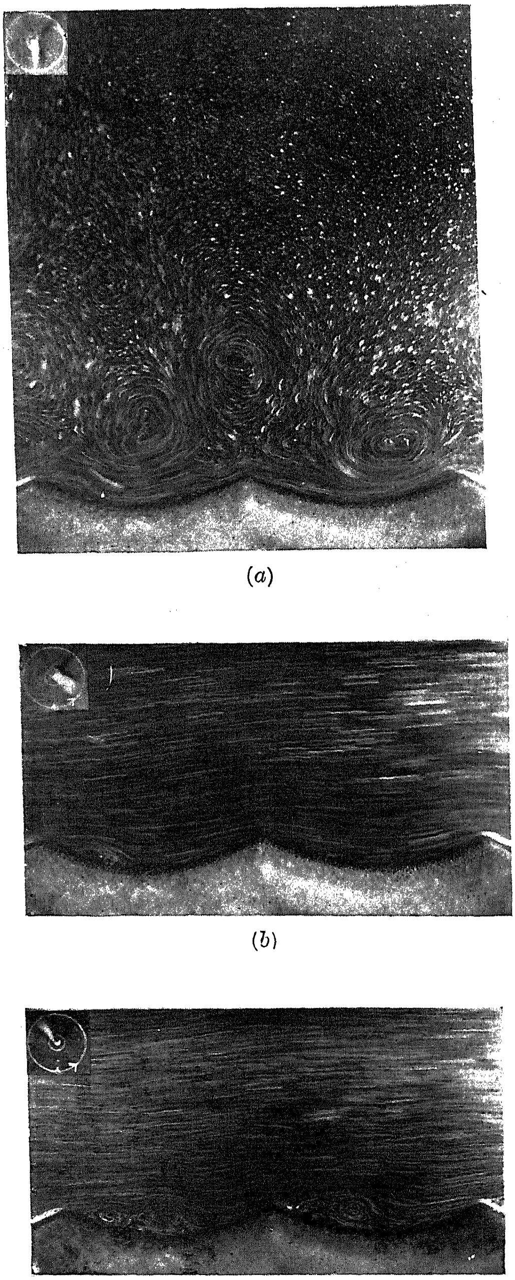


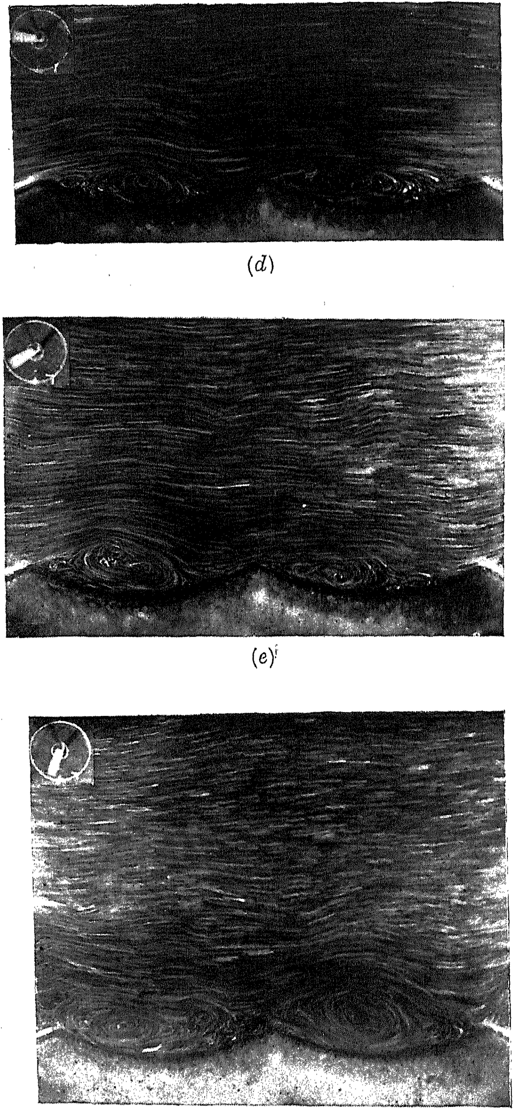
FIGURE 5 (*b*)



(*a*)

(*b*)

(*c*)



(*d*)

(*e*)

(*f*)

Direction of ripple	←	Ripple pitch	10 cm.
Half-period	3·4 sec.	Ripple height	1·5 cm.
Exposure	0·2 sec.	Stroke	20 cm.

FIGURE 6

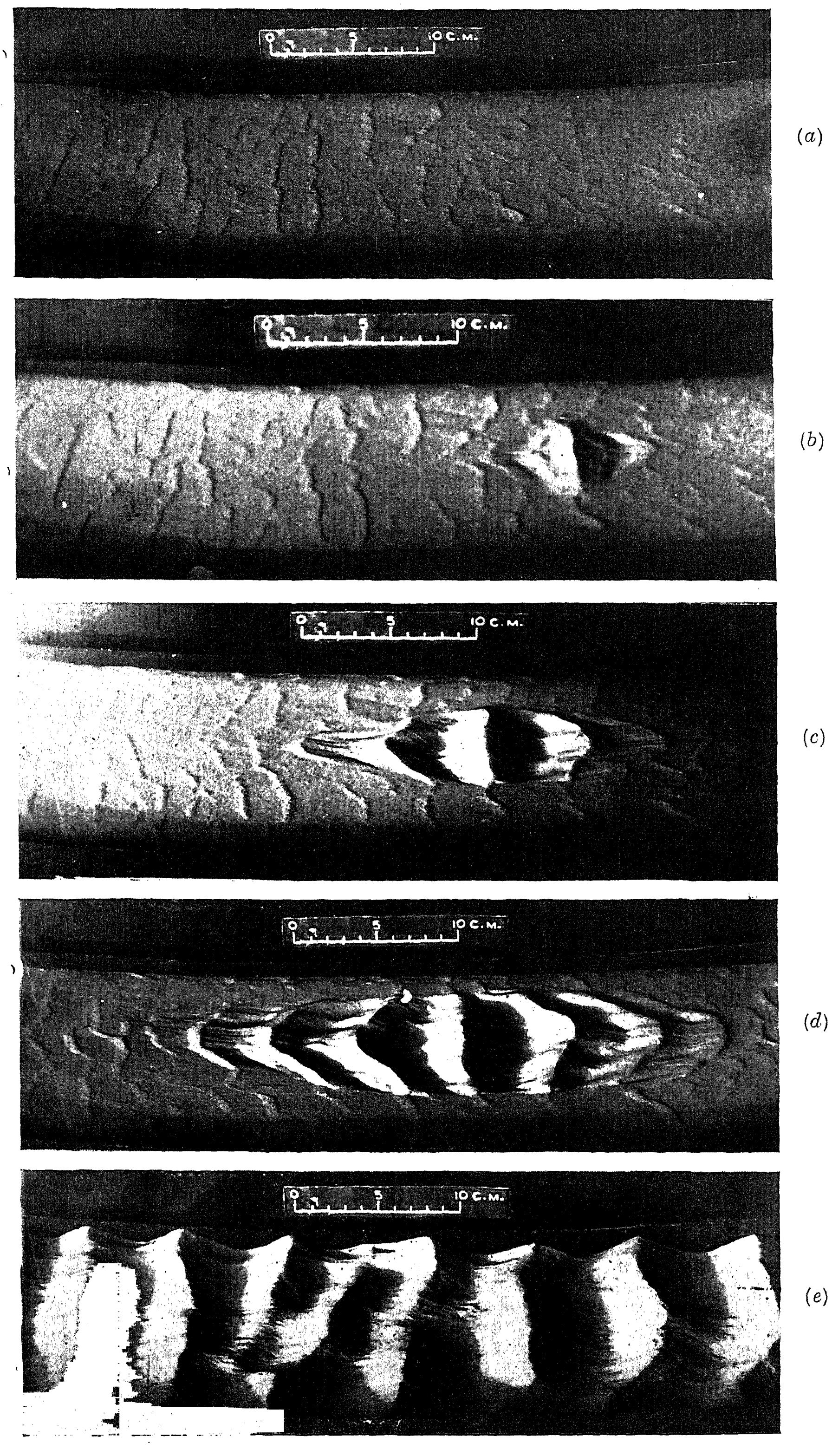


FIGURE 7

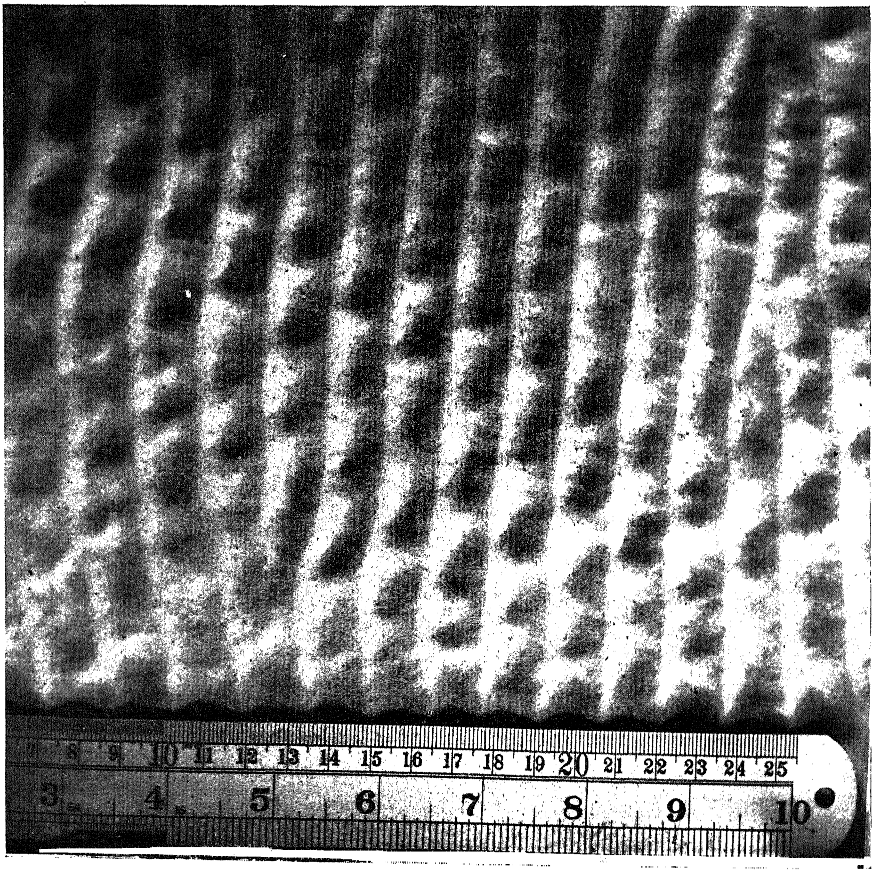


FIGURE 8

vortices must in these cases have been broken up into separate or nearly separate pieces, rather like strings of sausages.

The brick pattern appears in deep water on the sand bottom of a wave tank whenever the water amplitude is short enough; and the fact that it appears also under the very artificial experimental conditions where the bottom is oscillated through still water affords good evidence that no serious unrealities can be introduced by the experimental artifice. Further, the pattern persisted in the oscillating cradle even when the speed was raised to the theoretical breakdown value at which the end-of-stroke acceleration of the cradle must cause a general slipping of the sand grains, i.e. when $R\omega^2 = \mu g\dfrac{\sigma-\rho}{\sigma}$. This goes to reassure one as to the soundness of other experimental observations made close to this limit.

8. PHYSICAL LIMITS TO THE EXISTENCE OF VORTEX RIPPLES

Figure 2 gives a tentative schematic view of the various limits which appear from the experiments to bound the region (shown shaded) of water motion in terms of ω and R, in which the vortex-ripple mechanism can operate. Two examples have

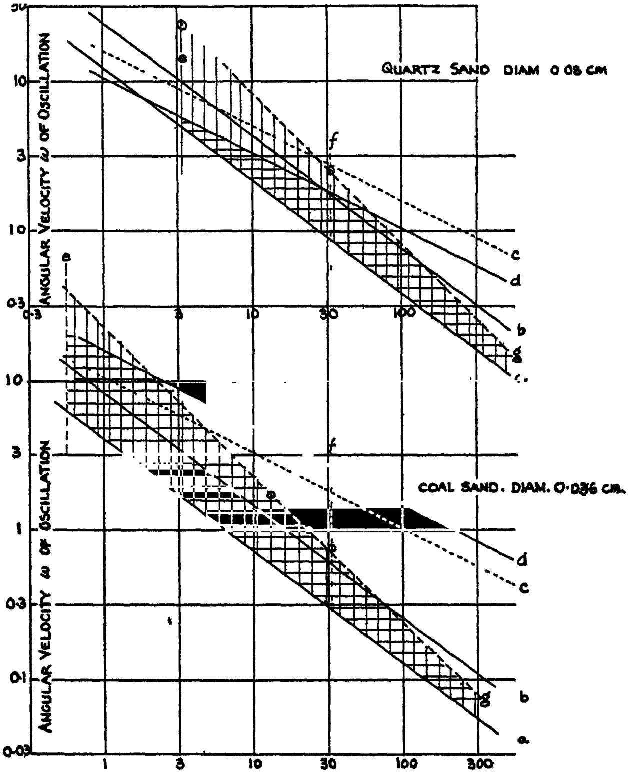


FIGURE 2

been taken: a quartz sand of 0·08 cm. diameter and a small grain coal sand of 0·036 cm. diameter. In the diagram each of the bounding lines shown is lettered as below. The limits for which there is direct evidence, either experimentally observed or calculated from wave theory, are shown as continuous lines. Inferred limits are shown as broken lines and limits set only by the apparatus used are dotted.

(*a*) Critical water motion for first disturbance of a smoothed-out sand surface. This was found experimentally, and is given by equation (4) in § 11, where it is further discussed.

(*a*′) Motion at which the vortex mechanism will continue to operate when once one ripple is formed. Speeds about 20 % lower than (*a*) (omitted from figure 2 for sake of clarity).

(*b*) Spontaneous formation of vortex ripples when no initial surface feature is present as a nucleus. Speeds about twice (*a*).

(*c*) Artificial limit set by experimental method, whereby the cradle acceleration $R\omega^2$ could not exceed $\mu g \frac{\sigma - \rho}{\sigma}$, where μ is the friction coefficient.

(*d*) Limit of bottom-water accelerations possible under stable travelling water waves. About $g/6$.

(*e*) Limit set to the smallest ripple pitch by size of the grain. This could only be observed for large grains. Ripples failed to form when ripple pitch was less than 30–50 grain diameters.

(*f*) Limit of stroke length; set by size of apparatus only.

(*g*) Observed failure of vortex mechanism at long strokes.

The lower speed limits set by (*a*′) or (*b*) will be seen to converge with (*d*) as the stroke amplitude decreases. Since both (*a*′) and (*b*) shift upwards with increase of either grain diameter or grain density, the region in which large heavy grains can ripple under free water waves is considerably restricted. But (*d*) can be raised under special conditions such as the violent oscillations which take place at the foot of a vertical wall against which waves are breaking. I obtained steep well-formed ripples by this means with pebbles 0·9 cm. in diameter (Bagnold 1940).

The limit (*g*) for long strokes is of some practical interest. Unfortunately, it lay in most cases beyond the experimental limit (*f*) of $R = 32$ cm. set by the apparatus, or else so close to the acceleration limit *c* as to be of doubtful significance. There was, however, fairly definite evidence that at high speeds the vortex mechanism began to fail in the case of the 0·08 cm. quartz sand before the stroke amplitude reached the experimental limit; and that failure occurred at shorter strokes as the grain size and/or density was reduced. The characteristic steep-sided vortex ripple, with its plume of whirling grains, gave place to a very flat streamline undulation over which a dense layer of moving grains glided low and smoothly like an oily liquid. There was also evidence of a change of pitch. In the case of the 0·036 cm. coal sand it rose abruptly from 11 to 15 cm.

Since the disappearance of the vortex occurs first at the highest oscillation speeds and longest strokes, it would seem to be a velocity effect; an inference which is supported by the three limiting values found for ω in the case of the fine coal sand of 0·036 cm. diameter, and plotted in figure 2*b*. If so, it can be represented by a 45° line in the diagram.

9. MAGNITUDE OF THE OSCILLATORY DRAG OF A RIPPLED SURFACE UNDER WAVES

In experiment 2, figure 5*b*, the mean drag τ per unit area of the artificially rippled plate was obtained from the measurement of the work done per oscillation in moving the plate through the water. For each of a range of different semi-amplitudes R, from 30 to 5 cm., records were made of the final steady angular speed ω maintained by each of a series of weights. The work done was given by the product of the effective weight (corrected for mechanical friction) and the distance it had fallen (the drum circumference). From this the mean drag τ throughout the oscillation was found by dividing by the total distance $4R$ travelled by the plate and by the area of the plate, both sides of which were rippled. In order to check for possible end-effects, the experiment was repeated with a short plate carrying one ripple only. The results of this check showed that end-effects were negligible.

The pitch/height ratio of the ripples was 6·7 to 1; and the ripple trough sections consisted of circular arcs meeting to form sharp crests at an angle of 120°. Two plates were used, on one of which the ripple pitch was 10 cm. and on the other twice the size.

The resulting figures for τ and ω for various values of R are plotted in figure 3. τ was found to be proportional to ω^2 for all values of R. But the variation of τ with R and with the ripple pitch p appeared to follow two different laws according to whether R was greater or less than p.

For all values of R/p less than unity, the drag coefficient $k = \tau/\rho\omega^2R^2$ was constant and independent of the ripple pitch, and thus

$$R/p < 1, \quad k = 0{\cdot}08. \tag{1}$$

Corroboration of this result seems to be forthcoming from steady-flow considerations. The drag, per plate, of a 'venetian blind' of a series of flat independent and widely spaced plates set at right angles to the direction of motion is $c\rho v^2A$, where A is the area of the plate and c depends on its length/breadth ratio. For infinitely long plates the value of c is given as 2 in *Smithsonian Tables*, Washington, 1934, table 159. Hence, if the breadth of each plate is $2h$, and their spacing is p, the drag coefficient per unit distance in the direction of motion is $2 \times 2h/p$. If now the plates are connected by a plane membrane through their central axes, each side becomes a surface roughened by transverse slats of height h, and the drag coefficient k for one side only will be $2h/p$. Assuming (i) that one can compare this with the r.m.s. velocity in the oscillating case, i.e. that it should be halved in order to become the k of equation (1) for which the maximum velocity $R^2\omega^2$ was used, and (ii) that the wide spacing in the steady flow case is comparable with the short stroke in the oscillating case, then, giving h/p its experimental value of 1·5/10,

$$k = 0{\cdot}15. \tag{1a}$$

This is of the same order as the experimental value, and is also independent of *R*. That it should be too big is to be expected, since the slats were vertical whereas the ripple troughs were filled in by smooth curves.

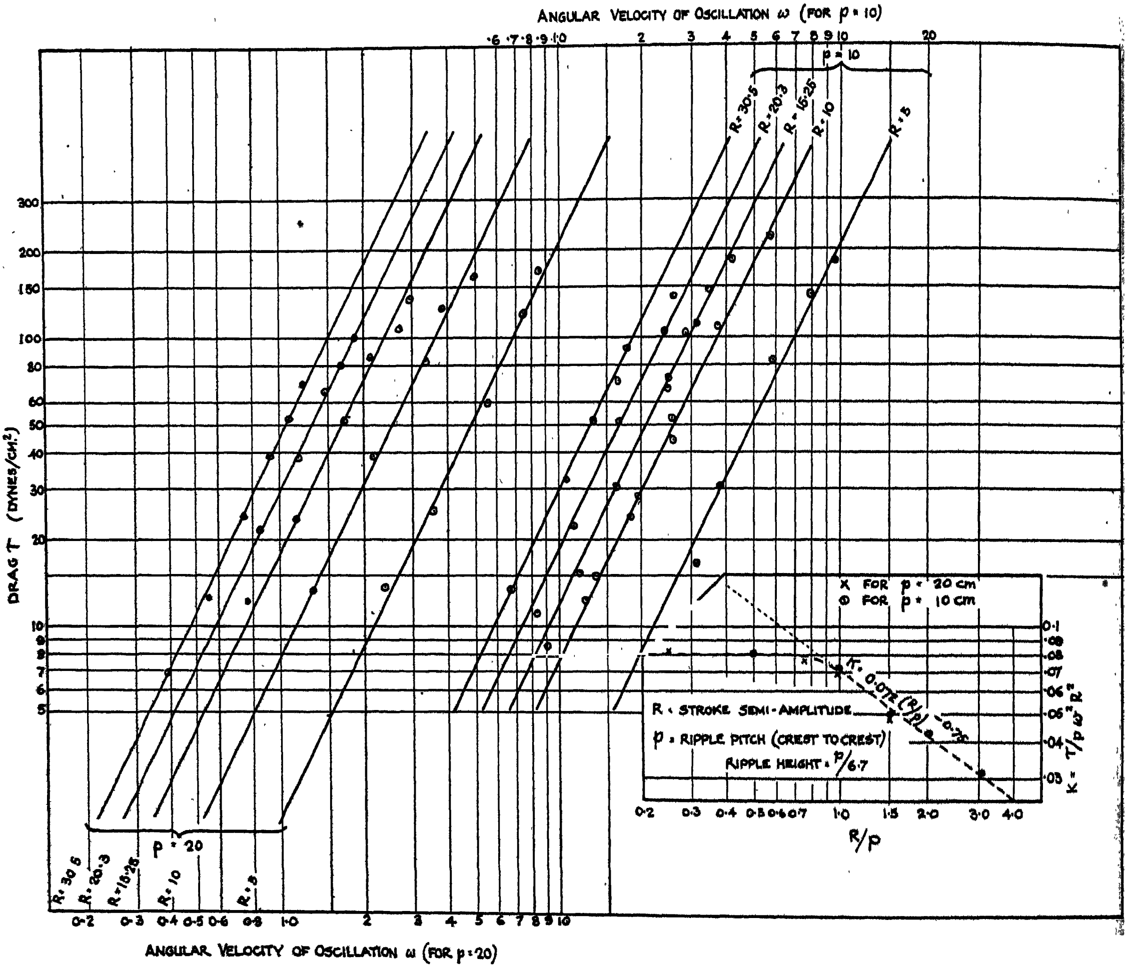


FIGURE 3

For values of *R*/*p* greater than unity, the experimental drag falls off with increasing values of *R* according to

$$R/p > 1, \quad k = 0{\cdot}072\left(\frac{R}{p}\right)^{-0{\cdot}75}, \tag{2}$$

though to what extent the numerical constant depends on the ratio *h*/*p* is not known.

In this case it is necessary to deal with the formation of a succession of transitory boundary layers, and confirmation of the results again comes from the steady flow case. Now I. Moyal, in an unpublished Ph.D. thesis,* found that the drag coefficient of a rough plate of length L moving longitudinally through the water is given by

---

* University of London, 1935.

$$1/\sqrt{k_1} = 4\cdot65 \log\left(\frac{2L}{d} \times 1\cdot64\right), \tag{3}$$

where d is the diameter of the sand grains with which the surface was roughened; and H. Schlichting (1936) gives values of the sand diameter d which is equivalent to the heights of other sorts of roughness. For vertical transverse slats of height h spaced at $p = 7h$, d is about $14h$.

Assuming that for the length L of the plate may be substituted the amplitude $2R$ of the oscillation, and putting $d = 14 \times 0\cdot15 \times p$, equation (3) becomes

$$1/\sqrt{k} = 4\cdot65 \log 3\cdot15 \frac{R}{p}. \tag{3a}$$

For $R/p = 1$ this gives $k = k_1/2 = 0\cdot09$, which may be compared with the experimental value of 0·072. For $R/p = 3$ it gives $k = 0\cdot025$, which is rather smaller than the experimental value of 0·03 to be found from figure 3. There is, therefore, a general agreement as regards the values of k, and k does fall off as R increases, though the log relation yields in this region of R/p a rather steeper curve.

10. RELATIONS BETWEEN DRAG RESULTS AND DIMENSIONS OF ACTUAL SAND RIPPLES

In attempting to relate the above results to those of the preceding sections, it should be borne in mind that:

(*a*) The final curve of figure 3 applies only to ripples of constant pitch/height ratio $\phi = p/h$, whereas in the actual sand ripple ϕ increased as a direct function of R, of the order of $(R/p)^{0\cdot25}$ for all observed values of R/p greater than unity.

(*b*) The limiting value of ϕ which the sand ripple attains where R/p is reduced to unity is doubtless set by the almost constant angle of repose of the sand grains on the sides of the ripple crest, and by the curvature of the trough bottom which could hardly have a radius less than that of the vortex which sweeps it out.

(*c*) With two exceptions, probably created in both cases by the excessive sharpness of the grains, R/p does not appear in actual sands ever to be able to fall much below unity, so it is possible that equation (1) never operates, because of the limiting value of ϕ being reached.

Now the whole phenomena of the vortex ripple suggests that when $R/p > 1$ there is a conflict between the end-of-stroke retardation, which appears to be responsible for the vortex which heightens the ripple crest, and the mid-stroke velocity of flow which tends to flatten out the crest; and the ratio retardation/velocity is $1/R$. Since from (*a*) above it seems that the actual value of k varies much more nearly as p/R than $(p/R)^{0\cdot75}$, it looks as if the 'natural pitch' P is such that

$$k \propto P \times \frac{\text{retardation}}{\text{velocity}},$$

and as if for any given grain diameter and stroke length the sand so adjusts itself as to give the maximum possible drag allowed by (i) the ratio retardation/velocity, (ii) the grain diameter of the sand, and (iii) the limiting geometry of the ripple.

11. SMOOTH SAND SURFACES. THE CRITICAL WATER MOTION FOR FIRST DISTURBANCE OF THE GRAINS

The sand sample was spread over the cradle of experiment 1, figure 5*a*, and the surface carefully smoothed. The oscillation speed was then raised till grain movement began.

An outstanding feature in this experiment was the absence of any signs of turbulence in the water, even at distances above the surface comparable with one grain diameter. The skein of colour threaded through the water which was left by a particle of dye as it fell to the bottom remained almost undisturbed after many oscillations had taken place.

For each of the sands used the values of ω corresponding to various semi-amplitudes R of the oscillation are plotted in figure 4, whence the critical angular speed ω appears to conform with the empirical relation

$$\omega = 21{\cdot}5R^{-0{\cdot}75}\gamma^{0{\cdot}5}d^{0{\cdot}325}, \tag{4}$$

where γ is the ratio $\dfrac{\sigma-\rho}{\rho}$ of apparent density, σ the density of the material having the values given in § 7, and ρ being unity for water.

In these measurements it is unlikely that the unreal conditions of still water and an oscillating cradle can have affected the results, since in this case the maximum accelerations of the moving cradle at the end of each stroke were very small compared with the limiting value of $\dfrac{\sigma-\rho}{\sigma}\mu g$; and, moreover, the initial grain movement took place at a later phase when the acceleration was considerably less than the maximum.

But there are several sources of possible error common to all measurements of the fluid velocity for first-grain movement:

(*a*) The difficulty of maintaining by eye, from one amplitude to another, and from one grain size to another, a consistent standard as to the proportion of moving grains on the surface which should constitute 'first movement'. This may well have introduced a systematic error which would affect the exponent of R, and would increase in magnitude as the stroke was shortened.

(*b*) The impossibility of using completely uniform material. The first surface grains to move might have been those of exceptional size or shape.

(*c*) In the case of the lightest material (coal) the possibility that the first grains to move might have had tiny air bubbles adhering to them. This error was watched for, but may not have been entirely excluded. Hence the actual value of γ in this case may have been rather too high.

With the justifiable introduction of g in association with γ in the form $(g\gamma)^n$, (4) may be written as

$$\omega = AR^{0{\cdot}75}(g\gamma)^{0{\cdot}5}\, d^{0{\cdot}25} d^{0{\cdot}075}, \tag{4a}$$

which is non-dimensional except for the final remnant exponent of d. But it seems likely from the apparent absence of turbulence that (4a) should also contain a factor involving the viscosity ν.

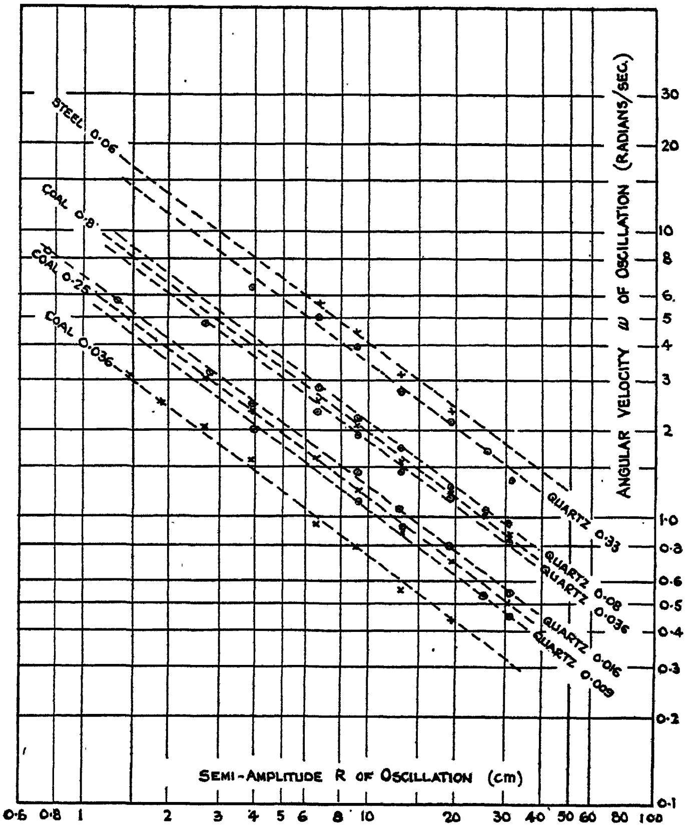


FIGURE 4

Further experiment is needed, using fluids of other viscosities, but (4) may be of some use as a first approximation in interpreting the results of model experiments with low-density artificial sands.

The above work was carried out on behalf of the Ministry of Supply, and my thanks are due to the Director-General of Scientific Research and Development for permission to publish.

REFERENCES

Bagnold 1940 *J. Instn Civ. Engrs*, **15**, 27.
McCowan 1891 *Phil. Mag.* **32**, 45.
Schlichting 1936 *Werft, R. Hafen*, pp. 99–102.

Note on R. A. Bagnold's empirical formula for the critical water motion corresponding with the first disturbance of grains on a flat surface

By Sir Geoffrey Taylor, F.R.S.

In §11 of his paper Bagnold calls attention to the fact that at the stage when oscillating water is just beginning to move sand grains no turbulence is produced even at a distance from the surface comparable with one grain diameter. The motion of water near a smooth plate which is oscillating horizontally has been calculated (Lamb 1932). In Bagnold's notation the horizontal velocity u of the water at height y above the plane is

$$u = \omega R e^{-\beta y} \cos(\omega t - \beta y), \tag{1}$$

where $\beta = \sqrt{(\omega/2\nu)}$ and ν is the kinematic viscosity.

At distances from the plate which are small compared with $1/\beta$ the motion is nearly the same as that of a liquid which is shearing uniformly at rate α, where

$$\alpha = \left[\frac{\partial u}{\partial y}\right]_{\eta=0} = \nu^{-\frac{1}{2}}\omega^{\frac{3}{2}} R \cos(\omega t + \tfrac{1}{4}\pi). \tag{2}$$

In fact, the motion through the whole layer which extends from the surface to $y = \sqrt{(2\nu/\omega)}$ is very nearly a uniform shearing flow, so that it seems justifiable to consider cases where grains have diameters less than $\sqrt{(2\nu/\omega)}$ as though they were being subjected to the stresses which would be applied to their surfaces by a uniform shearing motion in the fluid. In Bagnold's experiments ω varied from 0·5 to 6 sec.$^{-1}$, so that taking $\nu = 0{\cdot}011$ the above consideration might be expected to apply to grains whose diameters were below the range $\sqrt{\frac{2(0{\cdot}011)}{0{\cdot}5}} = 0{\cdot}2$ to $\sqrt{\frac{2(0{\cdot}011)}{6}} = 0{\cdot}06$ cm. Three grades of quartz sand used by Bagnold satisfied this condition, namely, those of mean diameter 0·0016, 0·016 and 0·036 cm. diameter.

The system of flow patterns produced in a uniformly shearing fluid by a grain of given shape when the diameter, d, the rate of shear α or the kinematic viscosity ν vary depends on a single variable only, namely the Reynolds number

$$z = \alpha d^2/\nu, \tag{3}$$

and the force which the fluid exerts on the grain must be of the form

$$F = d^2\rho(\alpha d)^2 f(z), \tag{4}$$

where f is a function which depends on the grain shape only and not on its dimensions. The condition that the grain will move must therefore be of the form

$$\rho\alpha^2 d^4 f(z) = A(\sigma - \rho)\, g d^3, \tag{5}$$

where A is a constant depending on the grain shape only. (5) may be rewritten in the form

$$\frac{\gamma g d^3}{\nu^2} = \frac{1}{A} z^2 f(z), \tag{6}$$

where

$$\gamma = \frac{\sigma - \rho}{\rho}. \tag{7}$$

(6) may now be regarded as an equation for (z) and can be solved to give

$$\frac{\alpha d^2}{\nu} = z = \phi(\zeta), \tag{8}$$

where

$$\zeta = \frac{\gamma g d^3}{\nu^2}. \tag{9}$$

In this form (8) may be compared with Bagnold's empirical equation (4) § 11 which gives the way in which ω depends on R, d and γ for substituting the maximum value of α from (z), namely $\alpha = \nu^{-\frac{1}{2}}\omega^{\frac{3}{2}} R$, (8) becomes

$$\omega = R^{-\frac{2}{3}}\nu d^{-\frac{4}{3}}\chi\left(\frac{\nu g d^3}{\nu^2}\right), \tag{10}$$

where $\chi(\zeta) = [\phi(\zeta)]^{\frac{2}{3}}$.

It will be seen that this analysis leads to the expectation that the exponent of R will be $-0{\cdot}66$ instead of the empirical value $-0{\cdot}75$.

The particular form of χ which most nearly coincides with Bagnold's empirical relation is $\phi(\zeta) = B\zeta^{\frac{5}{6}}$. This gives

$$\omega = (\text{constant})\, R^{-\frac{2}{3}}(g\gamma)^{\frac{5}{9}}\gamma^{-\frac{1}{9}} d^{\frac{1}{3}}, \tag{11}$$

making the exponent of $g\gamma$, 0·45 instead of the empirical 0·5; and that of d, $\frac{1}{3}$ instead of 0·325.

It is of interest to note the two limiting forms which might be expected when $\alpha d^2/\nu$ is very small or very large. In the former case (Stokes' law), $f(\alpha d^2/\nu)$ is proportional to $\nu/\alpha d^2$ and (10) reduces to

$$\omega = (\text{constant})\, R^{-\frac{2}{3}}(g\gamma)^{\frac{2}{3}}\nu^{-\frac{1}{3}} d^{\frac{2}{3}} \quad \left(\text{Stokes' law, small values of } \frac{\alpha d^2}{\nu}\right), \tag{12}$$

while in the latter

$$\omega = (\text{constant})\, R^{-\frac{2}{3}}(g\gamma)^{-\frac{1}{3}}\nu^{\frac{1}{3}} d^{-\frac{1}{3}} \quad \left(\text{constant drag coefficient, large values of } \frac{\alpha d^2}{\nu}\right). \tag{13}$$

It seems from (11) that the empirical formula gives exponents for γ and d which are intermediate between those of (12) and (13) but, so far as the variation of ω with d is concerned, the empirical formula is much closer to the Stokes' law limit (equation (12)).

It is of interest to make a rough approximate calculation of the critical value of α at which grains might be expected to begin to roll. Taking the case of a cubical grain of side d, if the tangential traction on the upper face due to viscous drag is the same as that on the surface on which the grain rests, it will give rise to a moment about the down-stream lower edge of the cube of amount $\mu\alpha d^3$. The face of the cube which faces the stream might be expected to experience a pressure of approximately the same magnitude as that in a half pitot tube placed facing the stream. This has been measured (Taylor 1938) and found to be approximately $1{\cdot}2\mu\alpha$ when the opening of the tube is so small that Stokes' law is obeyed. If the rear face of the cube has a suction of this magnitude, the moment of the normal components on the front and rear faces about the lower edge is $1{\cdot}2\mu\alpha d^3$. It seems likely that the normal component on the top face and the tangential components on the front and back faces will contribute less than the components already mentioned. The tangential drag on the side faces will make some positive contribution towards upsetting the cube, but it is difficult to make an estimate of its magnitude. These considerations lead to the conclusion that a cubical grain might be expected to roll when

$$2{\cdot}2\mu\alpha d^3 > \tfrac{1}{2}(\rho-\sigma)\,g d^4, \tag{14}$$

or expressed in terms of ω and R if

$$\omega > (4{\cdot}4)^{-\frac{2}{3}}(\gamma g)\,\tfrac{1}{3}\nu^{-\frac{1}{3}}R^{-\frac{2}{3}}d^{\frac{2}{3}}. \tag{15}$$

If this calculated value of ω is represented by the symbol ω_{cube}, the following table gives a comparison between ω_{cube} and the observed value of ω at which grains begin to move ($\omega_{\text{obs.}}$) taken from Bagnold's figure 4.

d (cm.)	R (cm.)	$\omega_{\text{obs.}}$ (sec.$^{-1}$)	ω_{cube} (sec.$^{-1}$)
0·009	11	1·0	2·0
0·016	8	1·5	3·63
0·036	10	1·8	5·4

It will be seen that $\omega_{\text{obs.}}$ and ω_{cube} are of the same order of magnitude, but that $\omega_{\text{obs.}}$ is greater than ω_{cube}. If similar rough arguments had been applied to the case of a hexagonal grain of height d in which the moment of the gravity component about the rolling edge is (volume of grain) $(\rho-\sigma)\,g\left(\dfrac{d}{2\sqrt{3}}\right)$ instead of $\left\{(\text{volume})\,(\rho-\sigma)\,g\left(\dfrac{d}{2}\right)\right\}$, better agreement between the calculated and observed values of ω would have been obtained.

REFERENCES

Lamb 1932 *Hydrodynamics*, 6th ed. p. 620. Camb. Univ. Press.
Taylor 1938 *Proc. Roy. Soc.* A, **166**, 479.

The photochemical polymerization of methyl vinyl ketone vapour

By T. T. Jones and H. W. Melville, F.R.S.

(*Received* 3 *August* 1945—*Read* 14 *February* 1946)

The photo-polymerization of methyl vinyl ketone vapour at pressures up to 60 mm. and temperatures between 0 and 60° C has been investigated with the following results. The kinetic order of the reaction increases with pressure and decreases with increasing temperature. The highest order recorded is 7·4. The temperature coefficient of the reaction is negative, but the rate does not obey the usual exponential relationship. The polymer produced is insoluble. These observations all point to the intervention of a branched chain mechanism, and on this basis the abnormal kinetics may be quantitatively explained.

The polymerization occurs by the splitting of the methyl vinyl ketone molecule into free radicals, probably $CH_3 + CH_2{=}CH.CO$. The former radical initiates a free radical polymerization which is terminated by a mutual interaction of two radicals in the vapour phase. The $CH_2{=}CH.CO$ radical appears to break up with the production of carbon monoxide, the $CH_2{=}CH$ radicals combining to yield butadiene which plays no further part in the reaction.

A micro-vapour pressure analyser is described which will deal with a few cubic mm. of vapour at N.T.P.

A fundamental problem in the initiation of polymerization reactions by means of ultra-violet radiation is the determination of the nature of the activated monomer molecule. On the one hand it may be that the activated monomer is a diradical and that molecules of monomer then add on to the free valencies so produced—the free radical mechanism—alternatively, the double bond may become excited in some special manner so that subsequent addition of monomer occurs with the preservation of an activated terminal double bond in the molecule. In previous papers an attempt has been made to discriminate between these two types by studying the kinetics of the free radical reaction by the intentional addition of hydrogen atoms or methyl radicals (Melville 1937). In the case of methyl methacrylate and vinyl acetate there is clear evidence for supposing that the direct photochemical reaction does not take place by the free radical mechanism, since the kinetics are so dissimilar to those of the reaction induced by free radicals.

The molecule of methyl vinyl ketone is of particular interest in this connexion because there is the possibility that, like many other ketones, it may break up into free radicals on absorbing radiation,

$$CH_3CO\,.\,CH{=}CH_2 = CH_3\overset{|}{C}O + \overset{|}{C}H{=}CH_2 \quad \text{or} \quad \overset{|}{C}H_3 + \overset{|}{C}OCH{=}CH_2.$$

Therefore the only possible kind of reaction would be of the free radical type initiated by one or other of the radicals so produced. In this paper an investigation has therefore been made of the kinetics in an attempt to establish the nature of the mechanism of polymerization. It so happens that many new peculiarities appear in this reaction which seem to be unique, and further emphasize the kinetic problems that still have to be solved in this branch of chemical kinetics.

EXPERIMENTAL

The general arrangement of the apparatus is shown in figure 1. The usual mercury-oil pump combination was used for evacuation. Owing to the solubility of methyl vinyl ketone and other vapours in the tap grease it was essential to use mercury cut-offs if the vapour had to be confined for long periods at constant pressure within a given part of the reaction system. The reaction system itself consisted of a wide-limb mercury manometer *GG*, a graduated gas burette *HH*, a U tube which could be immersed in liquid air, and a cylindrical quartz reaction vessel 6 cm. in diameter and 2 cm. thick. A side tube led to a special type of vapour analysis apparatus which will be described later. The movement of the mercury in the manometer was observed to within 0·05 mm. by a suitable travelling microscope. To avoid heating the reaction cell from the mercury lamp the former was immersed in a water thermostat which was fitted with a quartz lens and an acetic acid filter to cut off radiation at wavelengths less than 2500 A.

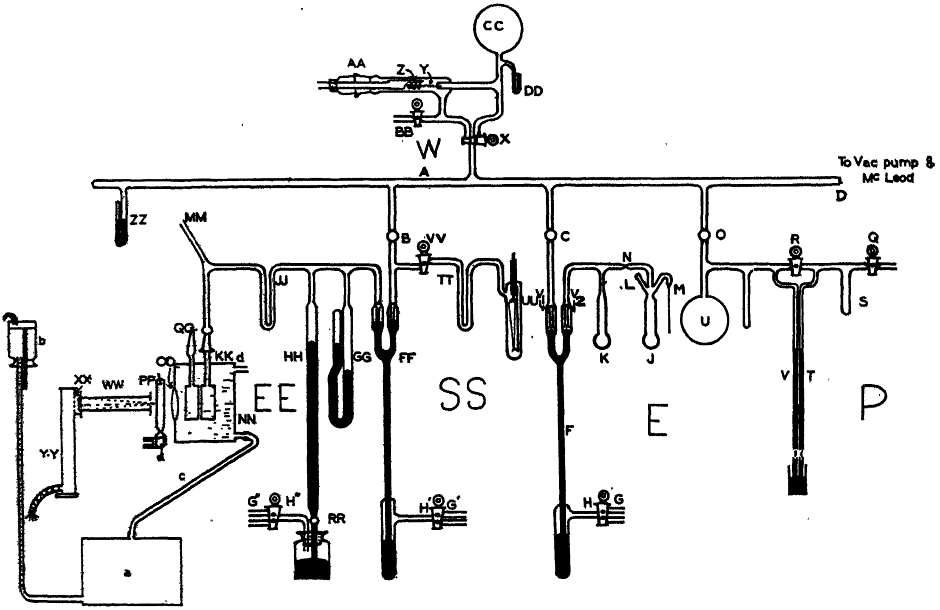


FIGURE 1. General arrangement of apparatus.

With this kind of arrangement it is therefore possible to follow the reaction at constant pressure or at constant volume. Small pressures of gas not condensable in liquid air are measured on the Pirani gauge. It was found necessary to protect this gauge by an additional liquid-air trap on account of the small amounts of monomer liberated from the tap grease and from the walls of the apparatus.

A check on the constancy of the output of the mercury lamp was kept by a photocell unit. The method employed was that suggested by Bowen (1936). This consisted essentially of a matrix of fluorescent crystals upon which the total radiation from the lamp fell, the fluorescent radiation being measured by a General Electric Company photocell electrometer valve combination unit. Using crystals of uranyl ammonium sulphate, $UO_2SO_4.(NH_4)_2SO_42H_2O$, Bowen has shown that from 2500 to 4500 A each quantum of absorbed radiation yields one quantum of fluorescent radiation at about 5000 A. Since the photoelectric sensitivity of a potassium cathode has a maximum at this wave-length such a cathode was used in the photocell. The detailed arrangement of the photocell end of the unit is shown in figure 2. *AB* is the fluorescent cell consisting of two thin polished silica plates 40 mm. in diameter enclosing the matrix of crystals embedded in purified paraffin wax. Beyond this cell is an Ilford γ filter to cut out all radiation shorter than the wave-length of the fluorescence before it impinges on the cathode *EF*. Since both sides of the cathode are sensitive a polished chromium-plated reflector is placed around the glass envelope of the unit. The whole system is enclosed within a brass cylinder and is connected by means of a multi-cored screened cable to the necessary control gear and potentiometer. The cell is operated in the usual way compensating for the change in the grid voltage on the electrometer valve when radiation falls on it by injection into the circuit of the necessary voltage from a potentiometer.

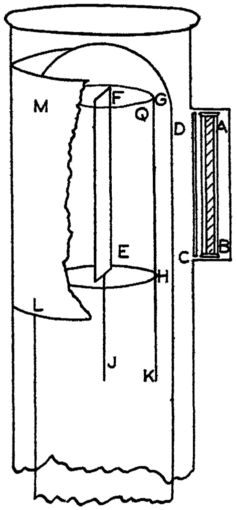


FIGURE 2. Photocell unit.

This form of fluorescent screen, while mechanically rigid, is not very efficient for a variety of reasons. Some experiments were made with two varieties of cell, *A* in which the tablet-like crystals were placed between the silica plates without any embedding agent, *B* in which the crystals were immersed in spectroscopically pure *cyclo*hexane. The sensitivities at 3000 A were in the ratios normal : *A* : *B* : : 1 : 1·59 : 3·0. (We are indebted to Mr Bowen for lending several cells of this kind for this and other experiments not recorded.)

The same system is also used for measuring the absorption coefficients of acetone and of methyl vinyl ketone. A variety of high-pressure mercury lamps was used for these experiments.

PREPARATION OF METHYL VINYL KETONE

2-Chloroethyl methyl ketone was prepared by absorbing ethylene in an ice-cold mixture of acetyl chloride and aluminium chloride. The resulting mixture was carefully poured into ice water to hydrolyse excess acetyl chloride. The chloroketone was then extracted with chloroform and the solution washed with water to remove acetic acid. The solution was dried with calcium chloride, the chloroform removed at reduced pressure, and the chloroketone distilled at 65° C. The latter was slowly added in 30 g. lots to dry dimethylaniline. At the reaction temperature the methyl vinyl ketone distilled quickly from the mixture. This was redistilled and collected at 77–80° C. The liquid is best stored in solid carbon dioxide but may be kept in a refrigerator at 0° C for 1–2 months. Further purification was effected by drawing the liquid into trap *J* in figure 1, which contained dry potassium carbonate. The middle fraction of the low-temperature distillation *in vacuo* was collected in another trap immersed in solid carbon dioxide.

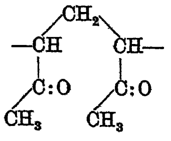
EFFECT OF PRESSURE AND OF TEMPERATURE ON POLYMERIZATION RATE

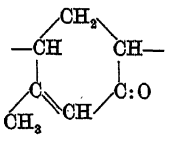
Methyl vinyl ketone vapour at pressures of 50 mm. and at 20° C polymerizes rapidly with the formation of a dense mist of solid polymer in the vapour phase on illumination with radiation between 2500 and 3500 A. While the predominant reaction is polymerization, prolonged illumination gives rise to small amounts of carbon monoxide, butadiene and traces of unidentified compounds. After several runs a thick skin of polymer forms on the surface of the reaction vessel. At the beginning of a series of runs with a clean reaction vessel there is a gradual increase in rate, for example,

number of run	1	3	5	7	9
rate in molecules per min. $\times 10^{-18}$	0·25	0·66	1·05	1·32	1·44

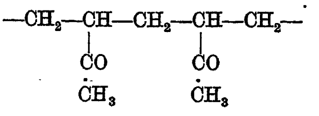
As will be shown later the reason for this increase is due to the fact that with uncovered silica walls the active polymeric molecules are rapidly deactivated on collision with the walls. As the walls become covered with polymer they become increasingly inert and a maximum rate is attained when all deactivation of active polymer occurs in the gas phase. If too much polymer is allowed to deposit the rate begins to fall owing to absorption and scattering of light by the film of polymer. These effects are particularly troublesome in working out reliably the kinetics of the reaction. In all this work therefore a reaction vessel had to be brought into an appropriate state in order to avoid wall effects and every experiment preceded and followed by one made under standard conditions.

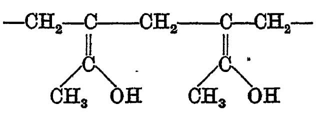
Nature of polymer. The photo-polymer produced in all these experiments was completely insoluble, thus suggesting a cross-linked structure. Further, the polymer finally became brownish in colour and exhibited a greenish hue in transmitted light. This coloration became more marked on warming the polymer *in vacuo* to about 100° C when a small quantity of colourless liquid, believed to be mainly water, was collected. Similar observations have been made by Marvel & Levesque (1938) on subjecting polymethyl vinyl ketone to pyrolysis. These authors attribute this to the dehydration of the polymer thus:

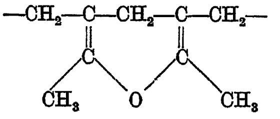


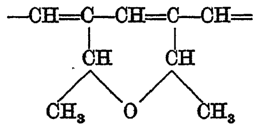


Another way in which the development of colour may be explained is the following set of rearrangements which can give rise to a kind of polyene structure known to be particularly strongly coloured (Bowen 1943).









It is unlikely that a system of conjugated bonds would extend very far for it would lead to much more intense coloration than is actually observed.

Effect of pressure. Preliminary experiments had clearly shown that the pressure dependence of the reaction rate was abnormal in the sense that not only was the order variable over a relatively small pressure range, but at high enough pressures became very high. The experimental procedure, therefore, was as follows. 30 mm. was fixed as a standard pressure and all the runs refer to the rate at this pressure, all experiments being carried out at constant pressure except where otherwise stated. For example, in order to obtain the increase in rate on raising the pressure to 40 mm. a clean vessel was coated with polymer until the rate had attained the maximum value. Next, runs were made with pressures of 30 and 40 mm. alternately until the diminution in rate at 30 mm. clearly showed that the polymer was absorbing an appreciable portion of the incident light. The mean value of the ratio of rates gave the information required. Table 1 gives a typical example of the results obtained with this procedure.

TABLE 1. TEMP. 15° C. CONSTANT-PRESSURE METHOD

pressure mm.	rate in burette units/min.	$\frac{\text{rate } 40}{\text{rate } 30}$	pressure mm.	rate in burette units/min.	$\frac{\text{rate } 40}{\text{rate } 30}$
30	12·9	—	40	25·5	3·2
40	42·0	4·0	30	10·5	3·2
30	14·6	4·0	40	25·25	3·9
40	45	4·7	30	6·9	4·1
30	10·9	4·3	40	17·4	—
				average	3·9

Ratio of rates here expressed as molecules/min.
1 burette unit at 30 mm. = $3{\cdot}4 \times 10^{16}$ molecules.

This series of experiments shows that the order of the reaction between 30 and 40 mm. has an extraordinarily high value of 4·8. The diminution in rate at 30 mm. towards the end of the series of experiments indicates the decrease in the intensity of the light absorbed by the vapour by a factor of 4. The ratio of rates is not noticeably affected. Such a high pressure dependence makes it extremely difficult to obtain a really accurate measure of the rate of reaction. This procedure was repeated for pressures of 10, 20 and 60 mm. The whole procedure was once more repeated for each of a series of temperatures. The results are summarized in table 2. In each series a clean vessel was used and several runs were made until the rate became steady.

TABLE 2

temp. ... 1° C		21° C		42° C		61° C	
pressure mm.	rate	pressure mm.	rate	pressure mm.	rate	pressure mm.	rate
20	$2{\cdot}0 \times 10^{19}$	30	$1{\cdot}19 \times 10^{19}$	32	$2{\cdot}80 \times 10^{18}$	30	$1{\cdot}05 \times 10^{18}$
10	$1{\cdot}7 \times 10^{18}$	20	$2{\cdot}30 \times 10^{18}$	16	$6{\cdot}7 \times 10^{17}$	60	$3{\cdot}80 \times 10^{18}$
20	$1{\cdot}66 \times 10^{19}$	30	$1{\cdot}23 \times 10^{19}$	32	$2{\cdot}65 \times 10^{18}$	30	$1{\cdot}32 \times 10^{18}$
10	$1{\cdot}33 \times 10^{18}$	40	$3{\cdot}6 \times 10^{19}$	46	$5{\cdot}95 \times 10^{18}$	60	$3{\cdot}86 \times 10^{18}$
		30	$1{\cdot}26 \times 10^{19}$	32	$2{\cdot}4 \times 10^{18}$	30	$1{\cdot}44 \times 10^{18}$
		10	$3{\cdot}43 \times 10^{17}$	60	$1{\cdot}36 \times 10^{19}$	10	$2{\cdot}71 \times 10^{17}$

Rate in molecules/min.

These results are shown in figure 3 by plotting log rate against log pressure, where it will be seen that at each temperature the order of the reaction is variable and increases with pressure over a narrow pressure range. Further, for any given pressure the order increases with decreasing temperature and attains its highest values at 1° C. The unfortunate situation arises that the low vapour pressure of the ketone prevents inquiry into a higher pressure region at low temperatures.

The abnormality of this polymerization is further shown when the log of the rate is plotted against the reciprocal of the absolute temperature for a series of pressures (figure 4). Again a curious behaviour is exhibited. Like other photo-polymerizations

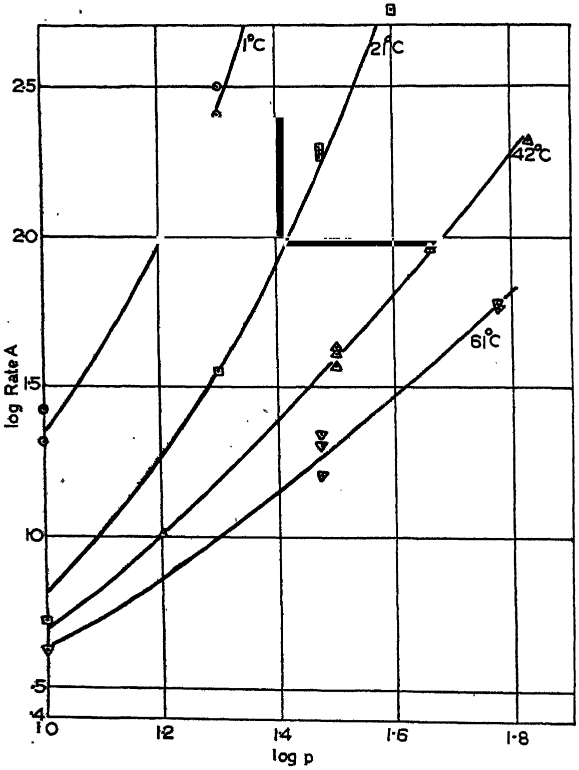


FIGURE 3. Variation of reaction rate with pressure.

in the gas phase the reaction possesses a negative temperature coefficient, but matters are additionally complicated by the fact that the lines are curved, therefore indicating a varying negative energy of activation. The dotted line indicates the limit of pressure which can usefully be employed at the temperatures indicated. The wide variation is shown more clearly by table 3 where the apparent energies of activation are computed from the curves given in figure 4. Similarly, the apparent orders for the reaction under various conditions are computed from the smoothed results given above and are shown in table 4.

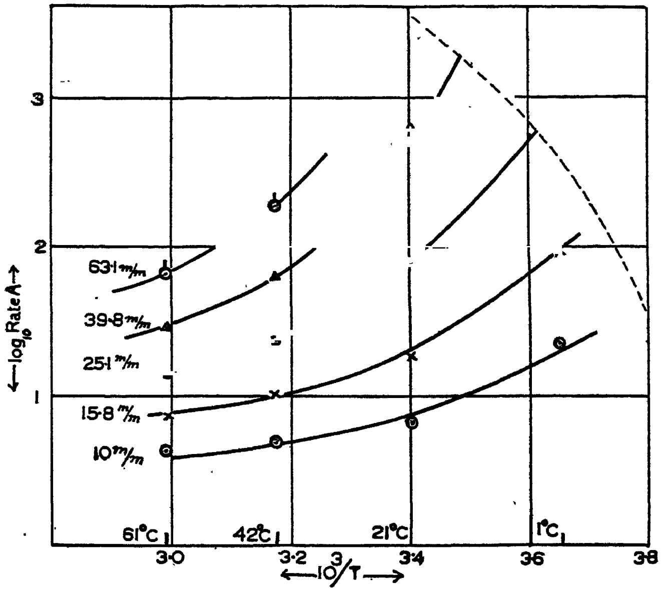


FIGURE 4. Variation of polymerization rate with temperature.

TABLE 3. APPARENT NEGATIVE ENERGIES OF ACTIVATION IN CAL. FOR PHOTO-POLYMERIZATION

pressure	temp.			
mm.	60° C	40° C	20° C	0° C
10	1,150	2,400	6,000	9,200
15·8	2,500	4,500	10,200	15,200
25·1	5,000	7,300	14,900	24,800
39·8	6,200	13,100	27,800	—
63·1	8,600	17,200	—	—

TABLE 4. ORDER OF REACTION

pressure	temp.			
mm.	60° C	40° C	20° C	0° C
10	1·19	1·60	2·05	3·06
15·8	1·38	1·86	2·64	3·98
25·1	1·50	2·16	3·21	7·40
39·8	1·67	2·38	5·67	—
63·1	1·88	2·53	—	—

The abnormal temperature and pressure dependence in this photo-polymerization is so far unique and must therefore be due to a factor not hitherto found in gas-phase polymerization. Of all the photo-polymerizations studied, namely, those of methyl methacrylate, methyl acrylate, vinyl acetate, methyl isopropenyl ketone and chloroprene, none except methyl vinyl ketone has given rise to completely insoluble polymers. It would therefore appear that the abnormal kinetics and cross-linking are interconnected in some way. The high apparent order of the reaction is reminiscent of the kind of behaviour obtained in the neighbourhood of the explosion limits of those thermal chain reactions whose velocity can be measured conveniently outside the explosion limits. In this kind of explosive reaction there is no question that the deviation and rapid increase in rate is due to the branching of the reaction chains giving rise to additional amounts of chain centres without actually causing explosion. By analogy then the pressure dependence of the methyl vinyl ketone polymerization might reasonably be explained. Abnormal temperature coefficients are also observed near explosion limits—a small increase in temperature leading to a very large increase in rate. With polymerization reactions normally possessing a negative temperature coefficient the phenomenon will be inverted for the lower the temperature the longer is the chain length and therefore the probability of the chain branching. The vapour pressure of the ketone unfortunately prevents access to what would appear to be an explosion region at low enough temperatures and high enough pressures, but the analogy is so close when coupled with the cross-linked nature of the polymer that there can be little doubt of the general formal similarity between combustion and polymerization in regard to branching.

Intensity exponent. This discussion of the above data does not permit of the formulation of a reaction mechanism. It is therefore necessary to look further into the kinetics to see whether the mechanism can be rigidly formulated, and the variation of rate with intensity gives useful information in this connexion. It has already been mentioned that starting with a clean reaction vessel the rate gradually increases. This can only be attributed to the fact that the active polymer is deactivated on collision with a silica wall, but is reflected without being deactivated from a wall covered with polymer. This kind of behaviour is often exhibited by free atoms and free radicals, and it is thus a pointer indicating that the active polymer may, in fact, be a free radical. In this circumstance it might be expected that the rate at which such free radicals combine or are otherwise destroyed on the walls is limited by the rate at which the radical diffuses to the wall. This will be proportional to the concentration in the gas phase and it is then easy to show that the rate of polymerization should be proportional to the first power of the intensity. As the walls become covered with polymer a greater fraction of the radicals will combine in the gas phase. Since this process necessarily involves two radicals the rate will be proportional to the square root of the intensity. The results in table 5 clearly show how the intensity exponent diminishes as gas phase combination becomes more important, thus supporting the above hypothesis.

TABLE 5

temp. ° C	condition of wall	pressure mm.	relative intensity	rate in mm./min.	intensity exponent
20	clean	30	0·384	0·61	1·0
20	clean	30	1·00	1·58	
20	partly covered	20	0·125	0·05	0·9
20	partly covered	20	1·00	0·32	
60	partly covered	23·5	1·00	0·025	0·7
60	partly covered	23·5	0·125	0·0058	
20	well covered	20	0·125	0·136	0·52
20	well covered	20	1·00	0·402	
20	well covered	20	1·00	0·384	0·49
20	well covered	20	0·125	0·138	

ACETONE SENSITIZED REACTION

It is now well established that the photochemical decomposition of acetone by radiation at about 3000 A involves the primary production of CH_3CO and CH_3 radicals. Further such radicals can initiate a polymerization reaction in which the active polymer is a large free radical. It was therefore important to attempt the analysis of the radical sensitized polymerization of methyl vinyl ketone in spite of the extreme difficulty that both ketones absorb at similar wave-lengths and therefore the two types of reaction cannot easily be separated and their individual kinetics established.

The first stage in carrying out this part of the work is the measurement of the extinction coefficients of acetone and of methyl vinyl ketone at a mean wave-length of 3000 A, namely the light transmitted through a 10 cm. silica bulb containing a solution of 145 g. of $NiSO_4.6H_2O + 41{\cdot}5$ g. of $CuSO_4.7H_2O$ in 1 l. of water (Bowen 1935). Owing to the weak absorption of both ketones and to the quick period variation in the output from mercury lamps, special arrangements were made to measure the absorption. In figure 5, $BXYCDE$ is the mercury lamp circuit with controlling resistance C and stabilizing choke D. XY is a resistance such that the fall of potential across it is about 2 V. From a small potentiometer FHJ an opposing potential is applied across XY so that there is no current through KL as indicated by the ammeter A. A galvanometer whose sensitivity can easily be varied by H, N and M is put across KL. This galvanometer thus records the short period fluctuations of current in the lamp circuit. The spot of light from the galvanometer in the anode circuit of the electrometer valve is projected on to the same scale as that used in the lamp circuit. To a first approximation it is assumed that the intensity of the light varies linearly with the current and hence the value of N was so adjusted that the spots of light moved together over the scale whenever there was a fluctuation of current in the lamp circuit. With the photocell covered the photocell circuit was balanced; on illuminating the cell containing monomer at the desired pressure the balancing voltage on the grid of the electrometer valve was adjusted to bring the spots of light into coincidence. Liquid air was then applied to a trap to remove the vapour and the balancing voltage redetermined. The difference of voltages divided by the voltage applied when the cell was evacuated is thus the fraction of light absorbed by the vapour. Results for acetone and for methyl vinyl ketone are given in table 6. The values of ϵ for acetone show only a slight tendency to increase at low pressures. This does not hold with methyl vinyl ketone which exhibits a marked and systematic decrease in ϵ as the pressure is reduced. The difficulty here is that at high pressures it is difficult to prevent polymerization, as unfortunately no inhibitor has been found effective for this purpose. The lower values are therefore more likely to be correct. It is not surprising that the extinction coefficient of methyl vinyl ketone is greater than that of acetone on account of the existence of the conjugated system $CH_2{=}CH.\overset{|}{C}{=}O$ and in fact it is surprising the increase is so small.

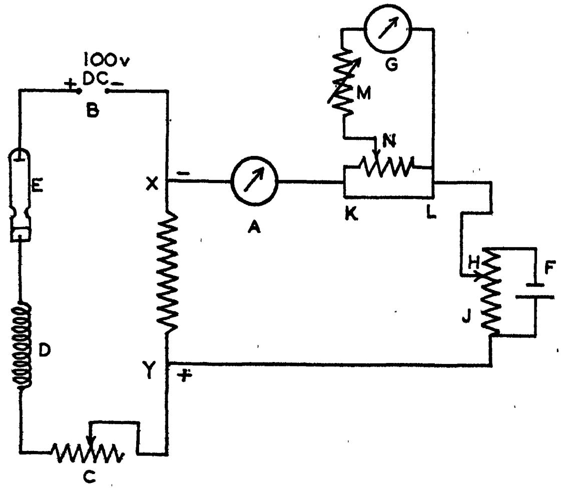


Figure 5. Compensating circuit for radiation from mercury lamp.

Table 6. Temperature 295° K. Thickness of absorbing layer 5 cm.

Acetone			Methyl vinyl ketone		
pressure mm.	percentage light transmitted	ϵ, molar decadic extinction coefficient	pressure mm.	percentage light transmitted	ϵ, molar decadic extinction coefficient
138	68·6	3·81	82	68·5	7·37
140	71·6	3·81	80	70·2	7·07
120	75·0	3·82	77	71·6	6·95
90	80·1	3·95	60	79·1	6·24
70	84·0	3·97	45	84·8	5·87
31	92·4	4·05	29	91·1	5·15
34	91·9	3·98			

That photo-decomposing acetone accelerates the polymerization is shown clearly by the results of a typical series of runs in table 7. The only way of getting a good comparison is to make alternate runs with and without acetone. As the polymer accumulated its extinction coefficient was measured and the rates observed and corrected to constant light intensity. The results are not regular enough to deduce the precise effect of acetone pressure but it is evident that pressures of acetone comparable with that of the ketone make a contribution to polymerization rate to an extent expected if both ketone and acetone decompose to give free radicals which start off polymerization.

TABLE 7. ACETONE-SENSITIZED POLYMERIZATION OF METHYL VINYL KETONE AT 60° C. METHYL VINYL KETONE PRESSURE 22 CM. (CONSTANT VOLUME METHOD)

acetone pressure (mm.)	19	0	10	20	0
corrected rate (mm./min.)	0·096	0·020	0·120	0·169	0·148
acetone pressure (mm.)	5	9·5	15·5	20	0
corrected rate (mm./min.)	0·146	0·199	0·228	0·267	0·22

Chain length of the acetone-sensitized reaction. In order to obtain some idea of the chain length in the acetone sensitized reaction, a long series of experiments was made with and without acetone until steady conditions were attained. The following figures show that part of the run when conditions were stable with a methyl vinyl ketone pressure of 22 mm. and temperature 50° C. These results give a mean value of 1·4 for the ratio with and without acetone. Independent measurements have shown that with acetone alone each carbon monoxide molecule produced in the photo-decomposition corresponds to the production of two methyl radicals (Melville & Tuckett, unpublished experiment). Although these measurements were made at 30° C the relation is almost certain to hold at 50° C where the quantum yield for the acetone decomposition is *c.* 0·6 (Leighton & Noyes 1941). Under the above conditions with 22 mm. acetone free radicals were produced at the rate of 0·0038 mm./min. Suppose that the methyl vinyl ketone polymerization occurs by the free radical mechanism and let x mm./min. be the rate of production of free radicals from methyl vinyl ketone at a pressure of 22 mm., then on the addition of 22 mm. of acetone the total rate of free radical production will be $x + 0{\cdot}0038$ mm./min. The rate of polymerization is proportional to the square root of the starting of chains and hence the following relation holds

acetone pressure (mm.)	0	21	0	23	0	23
corrected rate (mm./min.)	0·172	0·253	0·206	0·322	0·264	0·355

$$\frac{\text{rate of acetone-sensitized} + \text{M.V.K. reaction}}{\text{rate of M.V.K. reaction}} = \left\{\frac{x + 0{\cdot}0038}{x}\right\}^{\frac{1}{2}}.$$

From the three sets of results above, the values of x so obtained are 0·0039, 0·0043 and 0·006 mm./min. The ratio of methyl vinyl ketone polymerization rate to x gives the chain lengths which are 50, 55 and 45 respectively. While the chain length for the acetone reaction will be similar in magnitude, it is not possible to assess accurately its value on account of the concurrent reaction initiated in the first place by methyl vinyl ketone.

Rate of dissociation of methyl vinyl ketone to carbon monoxide. A series of runs were made to compare the rate of carbon monoxide production from methyl vinyl ketone and from acetone, since it was of interest to find whether the carbon monoxide production would give a measure of the rate of production of free radicals in the case of methyl vinyl ketone. On p. 29 we have seen that x has a value very close to that for acetone. If therefore each carbon monoxide molecule corresponds to the production of 2 free radicals, then the carbon monoxide production rate from methyl vinyl ketone should in this case be practically identical. Comparison of the rates of carbon monoxide production for methyl vinyl ketone and for acetone at 20 mm. and 50° C show that carbon monoxide production from acetone is 2·5 times that given by methyl vinyl ketone, thus indicating that apparently methyl vinyl ketone decomposition as reckoned by carbon monoxide production does not yield such a great number of free radicals as does acetone. The extinction coefficient of methyl vinyl ketone is certainly greater than that of acetone by a factor of 1·3 if ϵ-methyl vinyl ketone = 5·15 and hence the greater carbon monoxide production in the case of acetone must be due to the greater primary quantum efficiency of the dissociation reaction.

TABLE 8

				pressures in arbitrary units			
run	temp. °C	M.V.K. pressure mm.	acetone pressure mm.	CO obs.	CO expected from ketone	CO expected from acteone	total CO expected
A3	20	20	37·0	1·7	1·63	1·53	3·2
A5	20	20	11·5	2·1	1·94	0·58	2·52
H2	50	22	21·5	20·8	20·2	26·1	46·3
H4	50	22	24·0	19·3	18·6	21·2	39·8
B9	60·5	20	21·5	12·7	12·2	19·9	32·1
B11	60·7	20	57·0	14·5	12·3	53·0	65·3
D4	60	23	19·0	28·1	26·5	37·2	63·7

A number of experiments were then made with mixtures of methyl vinyl ketone and of acetone at different temperatures and intensities. The results are given in table 8. The most noteworthy feature is the fact that carbon monoxide production from the mixture is not numerically equal to the sum of the rates when the dissociation of the ketones is examined separately. The most obvious explanation would be that each ketone is acting as an internal filter for the other, but the fact of the matter as seen from the results on p. 28 shows that only 5 % of the light is absorbed by 20 mm. of acetone and 6 % by the methyl vinyl ketone. It will be further observed from an examination of table 8 that the observed carbon monoxide production is always slightly greater than that expected for the ketone alone, and that, if anything, high acetone pressures favour this slight excess. It is remarkable, for example, that with run B 11 nearly 50 units of carbon monoxide seems to disappear completely. At 60·7° C the acetone molecule undergoes immediate photo-dissociation so there can be no question of deactivation of excited molecules nor is it conceivable that the addition of say 20 mm. methyl vinyl ketone to 57 mm. acetone could possibly catalyse in some way the process of recombination of the molecular fragments. We therefore conclude that acetone initially is dissociated to $CH_3\overset{|}{C}O$ and CH_3 and that the $CH_3\overset{|}{C}O$ adds on to the methyl vinyl ketone molecule, thus

$$CH_3\overset{|}{C}O + CH_2{=}CH.COCH_3 \rightarrow CH_3.CO{-}CH_2{-}\overset{|}{C}H.COCH_3,$$

so effectively that it has no opportunity to dissociate to CH_3 and CO as happens to a large extent with acetone alone. These observations raise the question as to how the methyl vinyl ketone dissociates. Since carbon monoxide is produced there are three possibilities:

$$CH_2{=}CH.CO.CH_3 \rightarrow CH_2{=}CH.CH_3 + CO, \tag{a}$$

$$CH_2{=}CH{-} + CH_3CO{-}, \tag{b}$$

$$CH_2{=}CH.CO{-} + CH_3{-}. \tag{c}$$

Any one or all of the radicals in (*b*) and (*c*) might start off a radical polymerization. In view of the high probability that in the ordinary photo-polymerization of methyl vinyl ketone the mechanism is of the free radical type and that therefore initial decomposition of the molecule is into free radicals which are the initiating catalyst, it is unlikely that much decomposition according to scheme (*a*) occurs. A choice must therefore be made between (*b*) and (*c*). In view of the experience with acetone it would seem that if dissociation (*b*) occurred the acetyl radical would immediately react with methyl vinyl ketone and that consequently no carbon monoxide at all would be produced. If (*c*) is the only possible mechanism, the fact that carbon monoxide must be produced would imply that the radical $CH_2{=}\overset{|}{C}H.CO$ will break up to $CO + CH_2{=}CH{-}$. Whether the vinyl radical will then add on to another molecule of methyl vinyl ketone is not at all certain.

Products of dissociation of methyl vinyl ketone. In order to decide some of the matters discussed above it appeared to be necessary to investigate in greater detail the nature of the products of dissociation besides carbon monoxide. Without further complicating the issue the following hydrocarbons might possibly be found, ethylene, propylene and butadiene. Accordingly a gas analysis apparatus, based on vapour pressure measurements and of the smallest possible volume, was constructed as shown in figure 6. This was constructed of 1·5 mm. diameter tubing and consisted of an entrance tube *ABC* to admit a sample of vapour into the analyser proper *DEFGHP* and an evacuated side tube *JKL* serving as a manometric limb to measure the pressure in the space *FGHP*. Any unwanted vapour in *FGHP* would be removed along *POQR* which was connected at *R* via a tap to the main vacuum line. The procedure in analysis was as follows. The system is evacuated and all the mercury pulled down into *S*. The mercury is then raised just above *N* and *L* by opening *A*. The mercury level is then allowed to rise to *P*. By opening *L* the mercury rises to the same height in limbs *BC*, *DEF* and *JKL*. If now vapour is introduced into the system from *AB* the difference in levels in *BC* and *DEF* and *JKL* will indicate the pressure it exerts. The vapour is then admitted to *DFGH* by lowering the mercury as far as *CD*. *DE* is made of tubing 5 mm. internal diameter to prevent mercury being driven straight into *GH*. The height *DE* is 60 mm., thus allowing analyses up to vapour pressures of this magnitude. In order to obtain a representative sample, the first sample trapped in *PFGH* can be removed along *POQR* and a second sample introduced. Having trapped the required sample in *PFGH* the mercury is raised above *E* to a series of levels, the pressure exerted being noted as a function of volume with the side limb *GH* in a bath at a suitable fixed temperature.

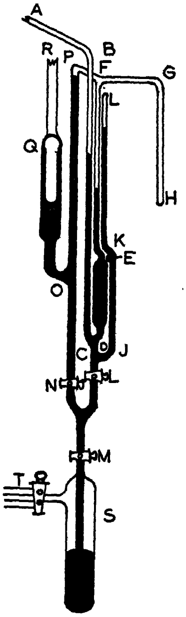


FIGURE 6. Micro vapour analyser.

During the course of the polymerization experiments it had been noticed that on prolonged illumination the pressure did not fall to zero. The vapour pressure of this residual material was somewhat greater than methyl vinyl ketone itself. In the analyser with liquid present, the pressure at −80° C was 13 mm. In another run illumination was continued first at a constant pressure of 54 mm. and then at constant volume until no further decrease could be detected. $2 \cdot 2 \times 10^{20}$ molecules were thus polymerized. 0·2 mm. of carbon monoxide was recorded. After removal of the carbon monoxide the whole mixture was condensed at −80° C, the pressure being 0·7 mm. and the most volatile constituents removed to the analyser by means of liquid air. The pressure exerted by the volatile constituents in presence of liquid was 12 mm. at −80° C. It may be noted that a sample of pure butadiene admitted to the analyser registered a pressure of 11·5 mm. at −80° C. A further sample from the reaction vessel was transferred to the analyser, cooled to −80° C and evacuated to remove what was presumed to be butadiene. The vapour pressure of the residual liquid was 60 mm. at 0° C compared with 20 mm. for methyl vinyl ketone. This liquid strongly coloured Schiff's reaction and smelled of acetaldehyde. In another run starting with an initial pressure of methyl vinyl ketone of 32 mm. the pressure fell after 330 min. illumination to a steady value of 6 mm., $8 \cdot 6 \times 10^{19}$ molecules being polymerized with the production of 0·5 mm. of carbon monoxide. The mixture was condensed in the reaction system at −80° C, the pressure being 0·7 mm. and the volatile constituents removed to the analyser. Again at −80° C the pressure registered was 13 mm. though this increased on considerable compression to 19 mm. which observation might indicate the presence of propylene in very small amounts (at −80° C the vapour pressure of propylene is *c.* 100 mm.).

The existence of butadiene in amounts comparable with those of carbon monoxide would therefore suggest that primary dissociation of methyl vinyl ketone is into $CH_2{=}\overset{|}{C}H.CO + CH_3$ and further the radical $CH_2{=}\overset{|}{C}H.CO$ breaks up into $CH_2{=}\overset{|}{C}H$ and CO. It is surprising that the vinyl radical is not wholly removed by its adding on to methyl vinyl ketone but instead it combines with another of the same kind to give butadiene. There is absolutely no indication of material as volatile as ethane and hence it must be presumed that all the methyl radicals react predominantly with methyl vinyl ketone to start polymerization. One possible way in which acetaldehyde might be produced is by dehydrogenation of perhaps methyl vinyl ketone by the CH_3CO radical.

GENERAL DISCUSSION OF THE RESULTS

It remains now to attempt to construct a mechanism to account for the abnormal kinetics of this reaction, taking as a basis that the ketone is initially dissociated to radicals and presuming that the methyl radical is primarily responsible for initiating a free radical polymerization of the ketone. In constructing this scheme it has to be borne in mind that the chains are certainly terminated by the mutual destruction in the gas phase of the polymer free radicals. It would appear that this destruction is brought about by combination rather than by disproportionation on account of the cross linked character of the polymer. In formulating the necessary kinetic expressions account must also be taken of some mechanism whereby kinetic branching can be introduced, but simultaneously the expression must be such that the rate is at all times strictly proportional to the square root of the intensity. In chain explosions where branching becomes at all noticeable the rate of initiation begins to have less direct control over reaction rate, and in the limit the intensity exponent should tend to zero. This diminution in intensity exponent has already been observed in the mercury sensitized reaction between hydrogen and nitrous oxide to which a small amount of oxygen has been added (Melville 1934). The simplest expression for the total active polymer concentration (P) is thus

$$-d(P)/dt = f(I) + \text{branching term} - k_t(P)^2 = 0,$$

where $f(I)$ is the rate of starting and k_t the termination coefficient. The simplest type of term for branching is of the form $k_B(P)^2(M)$. Inserting this term in the above equation, noting that even under the conditions of polymerization we may assume stationary state conditions (the main support for this assumption arises from the fact that the intensity exponent is strictly equal to 0·5 and as mentioned above had non-stationary conditions been approached this exponent would certainly have fallen below 0·5), hence

$$-d(M)/dt = k_p(P)(M) = k_p(M)\left\{\frac{f(I)}{k_t - k_p(M)}\right\}^{\frac{1}{2}}$$

$$= \frac{k_p(M)^{\frac{3}{2}}\,I^{\frac{1}{2}}}{\{k_t - k_p(M)\}^{\frac{1}{2}}}.$$

Since $I_{\text{abs.}} \sim (M)$ under experimental conditions at low pressures the order of the reaction will tend to 3/2. The results on p. 25 show that an order greater than 1 but less than 1·5 is obtained at the highest temperatures and lowest pressures. At high enough values of (M) the denominator in this expression will approach zero and thus very high rates and also very high orders will be obtained. For convenience the formula may be cast into the form

$$-d(M)/dt = A(M)^{\frac{3}{2}}/\{1-B(M)\}^{\frac{1}{2}}, \quad \text{where} \quad A = k_p(I/k_t)^{\frac{1}{2}} \quad \text{and} \quad B = k_b/k_t.$$

From the smoothed results given in table 2 we may set down for a given series of pressures how the rate varies with temperature. Using the values at two pressures, A and B may be computed at any one temperature and consequently the rate can be calculated at other pressures. Values of A and B have been computed using the data at the highest pressures so that an accurate value of B may be calculated. These calculated figures are shown alongside the observed values in table 9, where it will be seen that the agreement is as good as can be expected having regard to the simplified assumptions made in deducing the expression for the rate of reaction. Perhaps the important point to note is that abnormally high orders can easily be attained at pressures far below that at which the rate tends to indefinitely high values. Further, the type of mechanism proposed still permits the intensity exponent to remain at 0·5 in spite of the high degree of branching which occurs in the polymerization. If $\log A$ and $\log B$ are plotted against reciprocal absolute temperature straight lines are obtained corresponding to energies of $-6{\cdot}8$ kcal. for A and $-4{\cdot}8$ kcal. for B. This means that the energy of activation for branching is 4·8 kcal. less than that for termination.

TABLE 9

pressure	0° C rate		20° C rate		40° C rate		60° C rate	
mm.	obs.	calc.	obs.	calc.	obs.	calc.	obs.	calc.
10	20·9	13·4	8·0	15·6	4·97	7·7	4·17	3·53
15·8	105	—	22·4	34	10·5	15·9	7·68	7·2
25·1	1070	—	87·0	—	30·9	34·0	14·6	14·8
39·8	—	—	575	—	71·0	—	30·9	—
63·1	—	—	—	—	226	—	69·7	—

Another way of regarding the problem is to attempt to solve the expression

$$-d(M)/dt = k_p(M)\left\{\frac{f(I)}{k_t - k_b(M)}\right\}^{\frac{1}{2}}$$

by making use of the rates obtained for any one pressure at different temperatures. Thus if $A_p e^{-E_p/RT}$, $A_t e^{-E_t/RT}$ and $A_b e^{-E_b/RT}$ are substituted for k_p, k_t and k_b respectively then the expression assumes the form

$$\begin{aligned} -d(M)/dt &= A_p e^{-E_p/RT}(M)f(I)^{\frac{1}{2}}[A_t e^{-E_t/RT} - A_b e^{-E_b/RT}(M)]^{-\frac{1}{2}} \\ &= Ce^{-E_p/RT}[e^{-E_t/RT} - De^{-E_b/RT}]^{-\frac{1}{2}}, \end{aligned}$$

where $$C = \frac{A_p(M)f(I)^{\frac{1}{2}}}{A_t^{\frac{1}{2}}} \quad \text{and} \quad D = \frac{A_b}{A_t}(M)$$

i.e. $$\ln(\text{rate}) = \ln C - \frac{E_p}{RT} - \tfrac{1}{2}\ln\left[e^{-E_t/RT} - De^{-E_b/RT}\right].$$

There are five unknown quantities in this expression at any particular temperature and pressure, namely, C, D, E_p, E_t and E_b. If probable values are taken for E_p, however, some idea of the values of E_t and E_b can be obtained by using the experimental rates at different temperatures and constant pressure. The pressure chosen was 25·1 mm. and by using the rates at 0, 20 and 60° C the following values were obtained for E_t, E_b, $\ln C$ and D assuming $E_p = 2500$ and 4000 cal. The fifth column gives the calculated value of the rate at 40° C using the values found for E_t, E_b, $\ln C$ and D. This calculated rate should be compared with the experimental rate of 26·3.

$E_p = 2500$ cal.

E_t cal.	E_b cal.	$\ln C$	D	calculated rate at 40° C
20,500	13,050	−8·999	0·00000115	33·28
19,000	16,300	−8·089	0·007039	32·06
18,500	16,920	−7·902	0·05513	31·84
18,000	17,400	−7·936	0·3332	31·84

$E_p = 4000$ cal.

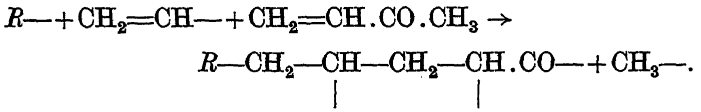
E_t cal.	E_b cal.	$\ln C$	D	calculated rate at 40° C
23,000	17,390	−8·663	0·00003376	32·51
21,720	19,680	−7·971	0·02365	32·06
21,000	20,700	−8·284	0·5758	31·92

The calculated rate approaches a minimum of about 31·9 for both values of E_p. An independent estimate of $\ln C$ or D cannot be made since the values of A_p, A_t and A_b are not known. It is interesting to note that the value of $E_t - 2E_p$ for those values of E_t and E_p which give rise to the minimum calculated rate, i.e. the rate nearest the experimental rate, is $18{,}000 - 5000 = 13{,}000$ cal. and $21{,}000 - 8000 = 13{,}000$ cal. for $E_p = 2500$ and 4000 cal. respectively. The value required previously for $E_t - 2E_p$ was 12,800 cal. The value of $E_t - E_b$ for those values of E_t and E_b which give rise to the minimum rate approaches zero however and this value should be around 5000 cal. to be in agreement with the previously estimated value of $E_t - E_b$. If $E_t - E_b$ is fixed at 5000 cal. then the following figures for E_t, E_b, $E_t - 2E_p$, ln C, D and calculated rate at 40° C are obtained:

E_p cal.	E_t cal.	E_b cal.	$E_t - 2E_p$ cal.	$\ln C$	D	Calculated rate at 40° C
2,500	19,800	14,800	14,800	−8·58	0·00013	32·56
4,000	22,800	17,800	14,800	−8·43	0·00016	32·40

The agreement between the values obtained for $E_t - 2E_p$ and the required value of about 12,800 cal. is still good. The above figures may therefore be taken to represent estimates of the extreme values of the quantities concerned. In the light of previous estimates for E_p higher values than 4000 cal. for a free radical mechanism are unlikely while lower values than $E_p = 2500$ seem to favour a greater divergence of the calculated rate at 40° C from the experimental value.

As is usual in the kinetics of complex reactions it is difficult chemically to formulate the kind of collision responsible for branching, though the above analysis definitely indicates the participation of two active or free radical polymers and a monomer, and as a result of this process a three-dimensional polymer must result. Tentatively we may suppose a reaction of the following type occurs. One of the free radicals is supposed to be the vinyl radical produced during the initial decomposition of methyl vinyl ketone. Although this radical is apparently not particularly reactive yet in the presence of another free radical, and one molecule of the monomer, the following reaction may occur, in which it will be seen that not only does the reaction give rise to a branched chain polymer but the number of free radicals is increased during this reaction:



$$R\text{—} + CH_2{=}CH\text{—} + CH_2{=}CH.CO.CH_3 \rightarrow R\text{—}CH_2\text{—}\underset{|}{CH}\text{—}CH_2\text{—}\underset{|}{CH}.CO\text{—} + CH_3\text{—}.$$

We should like to thank Bakelite Ltd. for a maintenance grant and the Department of Scientific and Industrial Research for a similar grant to one of us (T. T. J.). We also thank the Chemical Society for a grant for the purchase of silica apparatus. The experimental work described was done in the Colloid Science Laboratory, Cambridge, prior to the outbreak of war and we are very grateful to Professor E. K. Rideal for his encouragement and guidance throughout these researches.

REFERENCES

Bowen, E. J. 1935 *J. Chem. Soc.* p. 76.
Bowen, E. J. 1936 *Proc. Roy. Soc.* A, **154**, 349.
Bowen, E. J. 1943 *Ann. Rep. Chem. Soc.* p. 12.
Leighton, P. A. & Noyes, W. A. 1941 *Photochemistry of gases*, p. 359. New York: Reinhold.
Marvel, C. E. & Levesque, C. L. 1938 *J. Amer. Chem. Soc.* **60**, 280.
Melville, H. W. 1934 *Proc. Roy. Soc.* A, **146**, 737.
Melville, H. W. 1937 *Proc. Roy. Soc.* A, **163**, 511.

The free radical polymerization of the vapours of certain vinyl derivatives

By T. T. Jones and H. W. Melville, F.R.S.

(*Received* 3 *August* 1945—*Read* 14 *February* 1946)

The free radical polymerization of the vapours of vinyl chloride, acrylic nitrile, styrene, butadiene and methyl isopropenyl ketone has been investigated using methyl radicals from photo-decomposing acetone. Although all the reactions do not exhibit ideal behaviour it is possible by the application of kinetic methods to measure the ratio of the propagation to the termination coefficients for these polymerization reactions. In this way relative values of the quantitative tendency for a molecule to polymerize by a free radical mechanism can be computed.

An accurate measure of the tendency of a molecule to polymerize by way of a chain mechanism can only be gained by measuring the velocity coefficient for the interaction of the active polymer with the monomer. There may, however, be more than one type of active polymer, and therefore it will be necessary to specify its character when the appropriate coefficient is being computed. In a previous paper (Jones & Melville 1940) a method was suggested for making an approximate measurement of the magnitude of these bi-molecular coefficients for the vapour-phase polymerization of methyl acrylate. The method, however, was too crude to enable any conclusion to be drawn regarding the variation of these coefficients with molecular size. This method is essentially based on a knowledge of the primary photochemical efficiency of the polymerization, that is, the number of active monomers produced per quantum of absorbed light. If this quantity is not known the method becomes impracticable.

Many vinyl monomers, for example butadiene, acrylic nitrile and methyl isopropenyl ketone, do not exhibit direct photo-polymerization in the gas phase and hence the above-mentioned technique cannot be used. Under suitable conditions, for example, in aqueous emulsions, these molecules polymerize very readily, and it is therefore all the more necessary that some quantitative measure of their tendency to polymerize should be established. As will be shown in this paper they can be induced to polymerize by adding to the system free methyl radicals. At present, however, it has not been possible to devise a method to enable the calculation of the velocity coefficient of the growth of the active polymer to be made. On the other hand, in principle it is easy to obtain the ratio of this coefficient to that concerned in the termination of the growth of the polymer, namely, the combination of two free radicals. Pending the solution of the above-mentioned problem this ratio may be taken as a provisional measure of the tendency of the molecule to polymerize by the free radical mechanism. If later evidence indicates that the velocity of combination of free radicals of the type involved in polymerization is independent of their chemical nature, then the ratio is a direct measure of the propagation coefficient. Suppose the reaction is initiated by free radicals from acetone at a rate equal to $\alpha I_{\text{abs.}}$, where $I_{\text{abs.}}$ is the rate at which the radiation is absorbed by the acetone and α is quantum efficiency for the production of radicals. If the polymer radicals combine with each other it can readily be shown as a first approximation that the rate of polymerization is given by

$$R = k_p(\alpha I_{\text{abs.}})^{\frac{1}{2}}(M)/k_t^{\frac{1}{2}},$$

and the chain length ν by

$$\nu = \frac{k_p(M)}{k_t^{\frac{1}{2}}(\alpha I_{\text{abs.}})^{\frac{1}{2}}},$$

where k_p is the bi-molecular velocity coefficient for the growth of the polymer, k_t is the coefficient for the reaction of the radicals and (M) is the monomer concentration. It will be appreciated that only the ratio $k_p/k_t^{\frac{1}{2}}$ may be determined. In order to get some information about the magnitude of this ratio a study has been made of the free radical sensitized reactions of butadiene, styrene, acrylic nitrile and methyl isopropenyl ketone.

EXPERIMENTAL

The apparatus used was identical to that described in the previous paper. In using acetone as a source of free radicals a number of complications arise immediately, since the quantum yield for the decomposition of the acetone molecule is less than unity at temperatures below 75° C. This low coefficient may be due to the fact that excited acetone molecules may be deactivated without undergoing dissociation or it may be due to some recombination of the methyl and the acetyl radicals produced in the primary dissociation. The photo-chemistry of acetone has not yet been investigated in sufficient detail to permit of the calculation of the rate of production of acetyl and methyl radicals from the rate of absorption of radiation by the acetone molecule. The rate of production of carbon monoxide may, however, be used for this purpose, for it has been shown that under the conditions used in these experiments two free radicals are produced for each carbon monoxide molecule formed (Tuckett & Melville, unpublished experiments). This fact was established in the following manner. In the free radical polymerization of vinyl acetate, which may be induced by hydrogen atoms or methyl radicals, it is possible to adjust conditions so that the rates of these two reactions are identical. It is comparatively easy to measure the rate at which hydrogen atoms are incorporated in the polymer. It also happens that the rate of production of carbon monoxide from acetone in the absence of vinyl acetate is equal to twice the rate of production of hydrogen atoms, from which it may at once be concluded that two radicals are produced for each carbon monoxide molecule formed. Under the present experimental conditions the rate of production of free radicals is proportional to the pressure of acetone, since carbon monoxide production is directly proportional to the pressure of acetone, up to pressures of 80 mm. at least, as is shown in table 1. Hence in any particular acetone-sensitized polymerization the rate of production of chains will be proportional to the pressure of acetone, and if the intensity exponent is 0·5, i.e. the free polymer radicals mutually destroy each other, then the rate of polymerization should be proportional to the square root of the acetone pressure and to the first power of the monomer pressure, i.e.

$$-\frac{d(M)}{dt} = K(\text{acetone})^{\frac{1}{2}}\,(M).$$

TABLE 1. TEMPERATURE 60° C

run	pressure acetone (mm.)	rate of carbon monoxide production/mm./min., reduced to same light intensity (arbitrary units)
0·1	77	0·099
0·3	49	0·113
0·5	26	0·102

THE RADICAL-SENSITIZED POLYMERIZATION OF BUTADIENE

The results given in table 2 show that (*a*) the rate of polymerization is proportional to the square root of the acetone pressure for a constant pressure of butadiene for a constant intensity (figure 1); (*b*) the rate of polymerization tends to decrease slightly at higher butadiene pressures. The first observation is in accordance with the expectation and implies mutual reaction between the free radicals. The second result is, however, a little unexpected. Some additional observations on the rate of formation of carbon monoxide provide an explanation. The rate of carbon monoxide production in the presence of butadiene is always less than that obtained from acetone alone as shown in columns 5 and 6 of table 2. With pure acetone it is known that diacetyl is produced, presumably by the combination of free acetyl radicals initially formed by photo-dissociation. The majority of such radicals decompose to CH_3 and CO, but apparently in the presence of butadiene the CH_3CO radical may add on without dissociating further and only a fraction succeeds in breaking down to CH_3 and CO. The chain lengths, that is, the number of butadiene molecules polymerized per CH_3 radical produced, quoted in the last column are therefore calculated from the expected rate of production of carbon monoxide in the absence of butadiene. As a matter of interest it may be mentioned that the reduction in the carbon monoxide produced in the presence of monomer seems to be a characteristic of the polymerization of most vinyl monomers.

TABLE 2. TEMPERATURE 35° C

run	acetone pressure (mm.)	$\left(\begin{array}{c}\text{acetone}\\\text{pressure}\end{array}\right)^{\frac{1}{2}}$	butadiene pressure (mm.)	rate of production of CO (mm./min.)	expected rate of production of CO (mm./min.)	reciprocal rate of CO production	rate of polymerization (mm./min.)	chain length calculated from	
								CO produced	CO expected
J 1	21	4·6	0	0·00112	0·00112	893	0·0	0	0
J 2	21	4·6	31	0·00046	0·00112	2170	0·022	24·0	10·0
J 3	21	4·6	107	0·00026	0·00112	3800	0·014	27·0	6·3
J 4	21	4·6	52	0·00043	0·00112	2326	0·021	24·5	9·5
J 5	40	6·4	54	0·00058	0·00212	—	0·030	25·8	7·1
J 6	61	7·9	51	0·00078	0·00324	—	0·039	25·0	6·0

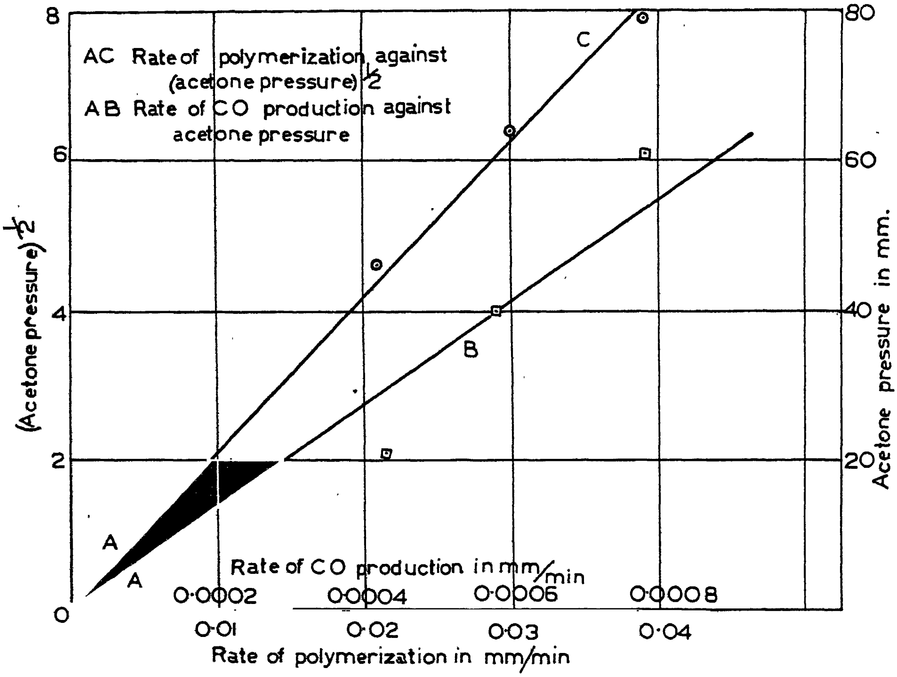


FIGURE 1. Rate of polymerization of butadiene as a function of acetone pressure.

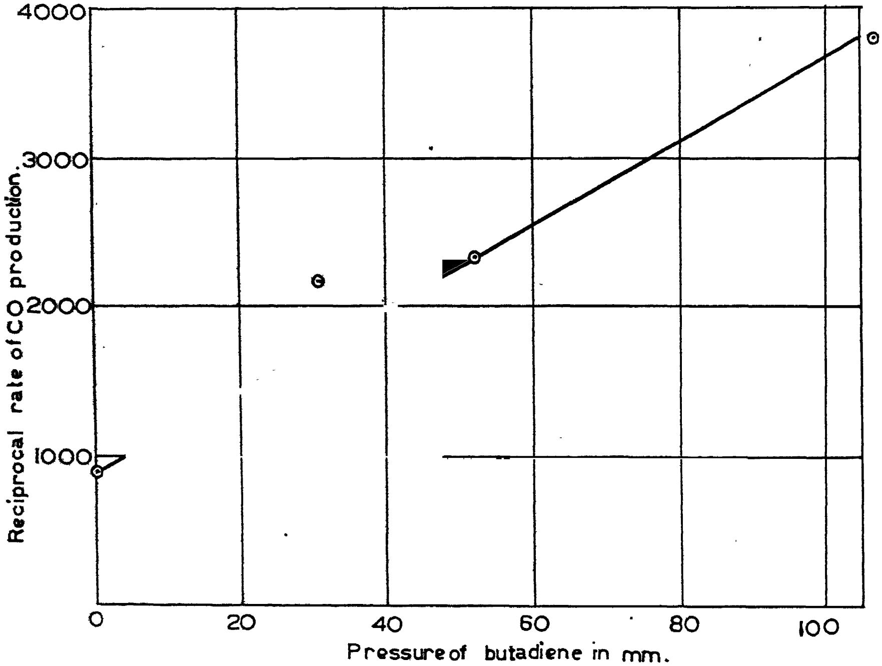


FIGURE 2. Rate of CO production from acetone in pressure of butadiene.

The reciprocal of the rate of production of carbon monoxide turns out to be proportional to the butadiene pressure (figure 2). This observation can be quantitatively accounted for by adopting the following mechanism. The concentration of the acetyl radicals (Ac) is defined by

$$d(\mathrm{Ac})/dt = \beta I_{\mathrm{abs.}}(A) - k_1(\mathrm{Ac})(B) - k_2(\mathrm{Ac}) - k_3(\mathrm{Ac})(A) = 0,$$

where $\beta I_{\mathrm{abs.}}$ is the rate of production of acetyl radicals per mm. of acetone, $k_1(\mathrm{Ac})\,B$ is the rate at which acetyl radicals combine with butadiene, $k_2(\mathrm{Ac})$ is the rate of spontaneous dissociation, and $k_3(\mathrm{Ac})(A)$ is their rate of dissociation in collisions with acetone molecules. This leads to

$$\frac{d(\mathrm{CO})}{dt} = \frac{k_2\beta(I_{\mathrm{abs.}})(A)}{k_1(B)+k_2+k_3(A)},$$

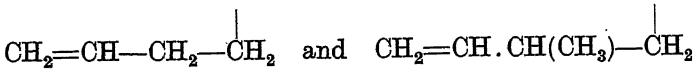
which is in agreement with the experiment in so far as carbon monoxide production as a function of butadiene pressure is concerned. In the absence of butadiene it has been pointed out that carbon monoxide production is accurately proportional to the acetone pressure and consequently it may be concluded that $k_2 \gg k_3(A)$. Further, if (B) is finite, similarly carbon monoxide production, though less than when $(B) = 0$, will also be proportional to acetone pressure. Figure 1 shows that this result holds only approximately. The conclusion therefore is that methyl or acetyl radicals lead to the free radical polymerization of butadiene and that the value of $k_p/k_t^{\frac{1}{2}} = 1{\cdot}5 \times 10^{-5}$ when the pressure of butadiene is given in mm.

HYDROGEN ATOM-CATALYSED POLYMERIZATION OF BUTADIENE

The results obtained in the hydrogen atom-sensitized reaction which are given in table 3 show quite conclusively that this reaction, curiously enough, does not lead to true polymerization of the butadiene at all. In the first place the total rate of decrease of pressure which may be taken as a measure of the rate of polymerization is nearly proportional to the first power of the intensity of the light used to make the hydrogen atoms. This is to be contrasted with the square root dependence on intensity obtained with methyl radicals. Furthermore, the apparent decrease in the pressure of butadiene is practically equal to that of the hydrogen used up. Therefore it is reasonable to assume that the butadiene is converted into a dimer by the action of the hydrogen atoms. The dimer, of course, is sufficiently volatile to contribute to the total pressure in the system. The simplest sequence of processes which may be put forward to explain these results is as follows:

$$\mathrm{H + CH_2{=}CH{-}CH{=}CH_2}$$
$$\downarrow$$
$$\mathrm{CH_2{=}CH{-}CH_2{-}CH_2{-}}$$
$$\downarrow$$
$$\mathrm{CH_2{=}CH{-}CH_2{-}CH_2{-}CH_2{-}CH_2{-}CH{=}CH_2}$$

two of the butadiene radicals combining. In this way one molecule of hydrogen disappears for two molecules of butadiene which give rise to an octadiene. Again, it is noteworthy that in the first stages of the methyl radical-sensitized reaction an essentially similar process occurs, but apparently the initially produced free radical may combine with many more ordinary butadiene molecules without the growth coming to such a premature stop as occurs with hydrogen. This would seem to imply that the two free radicals



$$CH_2{=}CH{-}CH_2{-}\overset{|}{C}H_2 \quad \text{and} \quad CH_2{=}CH.CH(CH_3){-}\overset{|}{C}H_2$$

possess very different reactivities towards butadiene molecules. The surprising fact is that it is the first radical that is apparently the less reactive towards butadiene in this respect. It is unnecessary to work out the quantitative aspects of this mechanism, since the hydrogen atoms are removed from the system by a reaction of the first order with respect to the concentration, and hence the total rate of reaction inevitably must be proportional to the first power of the radiation intensity.

TABLE 3. TEMPERATURE 35° C

run	butadiene pressure (mm.)	hydrogen pressure (mm.)	relative intensity	rate of reaction (mm./min.)	intensity exponent	total decrease in pressure (mm.)	decrease in hydrogen pressure (mm.) A	apparent decrease in butadiene pressure (mm.) B	ratio B/A
J 9	30	35	I_0	0·085	1·06	9·0	4·46	4·54	1·02
J 10	30	34	$0{\cdot}125I_0$	0·0095	1·27	3·8	2·05	1·75	0·85
J 11	31	33	$0{\cdot}384I_0$	0·0253	0·89	6·0	3·16	2·84	0·90

Further evidence that butadiene can be polymerized by free radicals has been obtained by interpolymerizing this substance with methyl vinyl ketone. As has been shown in the previous paper, methyl vinyl ketone polymerizes by a free radical mechanism and the free radicals were used to interpolymerize methyl vinyl ketone and butadiene. Prolonged runs with butadiene-methyl vinyl ketone mixtures were carried out with a view to finding if any butadiene was being used up in this manner.

Run H 14 (table 4), starting with 31 mm. methyl vinyl ketone and 24 mm. of butadiene, is typical of the few runs carried out; while run H 16, carried out with an initial pressure of 31 mm. of methyl vinyl ketone alone, serves as a comparison.

TABLE 4

run	initial pressure methyl vinyl ketone (mm.)	initial pressure butadiene (mm.)	estimated amount butadiene used up (mm.)	total amount polymerized (mm.)
G 10	31	25	4	16
H 14	31	24	3·5	16
K 5	31	31	0·8	4
K 9	31	35	0·6	8

Figure 3 shows the course of these runs. Curve *AB* gives the course of polymerization in presence of butadiene and curve *DEF* in its absence. Curve *DEF* is displaced a little to the right to distinguish it from curve *AB*.

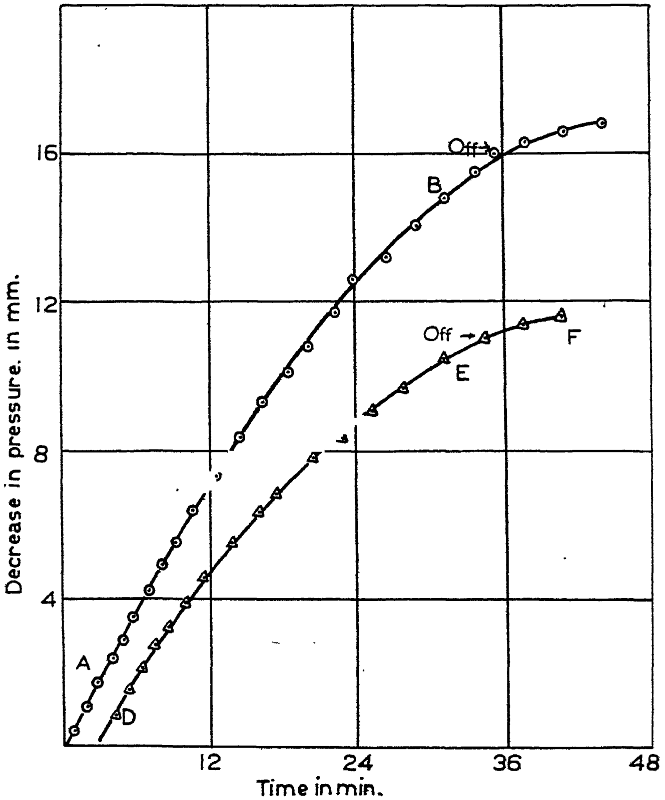
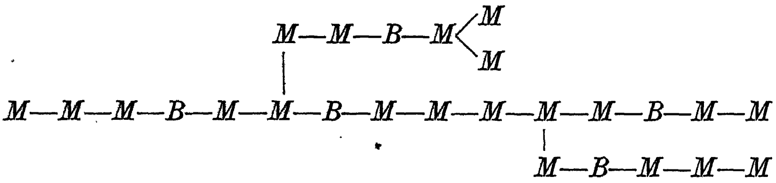


FIGURE 3. Polymerization of methyl vinyl ketone in presence of and in absence of butadiene. Curve *AB*, butadiene present.

Both polymerizations start off at equal rates of about 0·6 mm./min., but the direct polymerization is seen to fall off quicker with decreasing pressure than the polymerization in the presence of butadiene. This points to the fact that butadiene is used up, thus leaving more methyl vinyl ketone to maintain the rate. The rate at *B* after about 16 mm. of mixture had polymerized was about 0·29 mm./min., while the rate at *E* after about 10 mm. of methyl vinyl ketone had polymerized was 0·20 mm./min., showing how the normal polymerization has fallen off in comparison with that in presence of butadiene.

Further evidence that butadiene was used up was obtained by fractional distillation of the mixture before and after illumination. First the mixture was all condensed in the side tube of the analyser described elsewhere, by means of liquid air, before illumination. Liquid air was then applied to the reaction system and then solid carbon dioxide-acetone mixture to the analyser side tube. In this way mainly butadiene was fractionally distilled at −80° C into the reaction system where the pressure of the distillate at intervals of a few minutes was read. The same procedure was repeated with the mixture left at the end of a prolonged illumination, and it was found in every case that less distillate could now be obtained than was obtained before illumination, indicating of course that the difference, or a quantity proportional to the difference, had been used up. Table 4 gives the results obtained in this way.

Except for run K 9 it is seen that approximately one butadiene molecule is used up every three or four molecules of methyl vinyl ketone. If this meant that each butadiene molecule used up stopped a chain then the chain length would be extremely small and only of the order of three or four. However, since the rate of methyl vinyl ketone polymerization is maintained in presence of butadiene this probably means that true interpolymerization with butadiene takes place, the final polymer having a structure of the branched type similar to that of methyl vinyl ketone polymer, e.g.



M being methyl vinyl ketone and *B* butadiene units.

To obtain some comparison of the rate of acetone-sensitized butadiene polymerization and of the polymerization of methyl vinyl ketone in presence of butadiene, the runs in table 5 were carried out consecutively.

TABLE 5

run	pressure butadiene (mm.)	pressure methyl vinyl ketone (mm.)	pressure acetone (mm.)	rate of polymerization (mm./min.)	estimated rate of butadiene disappearance (mm./min.)
K 5	31	31	0	0·65	0·12
K 6	31	0	31	0·03	0·03

These results further show that since the rate of free radical production is approximately the same for acetone and methyl vinyl ketone, the increased rate of butadiene disappearance in the presence of methyl vinyl ketone is probably due to the presence of additional centres produced by branching.

POLYMERIZATION OF VINYL CHLORIDE

Vinyl chloride is very difficult to induce to polymerize in the gas phase, and the following two runs show the order of magnitude of the chain length calculated from the observed small rate of polymerization and the rate of production of free radicals.

TABLE 6. TEMPERATURE 20° C

pressure vinyl chloride (mm.)	pressure acetone (mm.)	rate of polymerization (mm./min.)	amount of CO produced (arbitrary units)	expected CO production	rate of free radical production (mm./min.)	chain length
31	32	0·005	232	180	0·003	1·5
156	31	0·02	217	142	0·003	7·0

THE DIRECT PHOTOPOLYMERIZATION OF VINYL CHLORIDE

Although vinyl chloride is transparent to light of wave-lengths 2500 A, on exposure to the light emitted from a zinc spark which contains a strong group of lines at about 2100 A polymerization takes place.

This polymerization again is not fast, and pressures of the order of 600 mm. must be used so as to get a measurable reaction. That polymerization does take place, however, is made evident by the formation of a cloud of polymer which soon settles on the sides of the reaction cell, and owing to its non-transparency begins to absorb the active radiation. It was found that the heat from the zinc spark produced temperature changes in the reaction system, while the reaction cell could not be thermostated in water since the latter completely absorbed the effective radiation. The exposures had thus to be carried out with the reaction vessel in air, and at the end of the run the vessel had to be cooled down to a suitable temperature and the decrease in pressure noted. This rather laborious method prevented any accurate kinetic measurements being made.

Figure 4 is a diagram of the apparatus. This consists of two parts: (*a*) the reaction system proper consisting of the reaction cell *B* with its associated liquid air trap and manometer, and (*b*) an ammonia photometer consisting of the quartz cell *A* immediately in front of *B*, together with the associated Pirani gauge and liquid air trap.

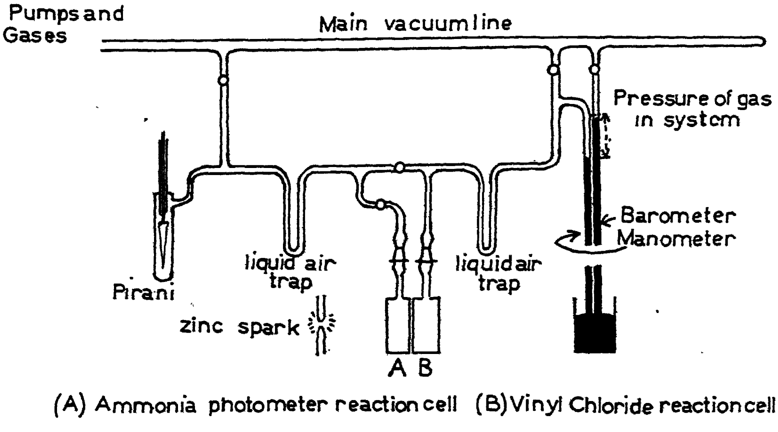


FIGURE 4. Apparatus for polymerization of vinyl chloride.

Figure 5 shows the housing of the 3000 W zinc spark. An alternating current of 6000 V was applied across the two zinc poles *P* and *P′* in the form of cylindrical zinc rods about 10 or 12 mm. in diameter. The poles were contained in the double-walled box *B*, well lagged with asbestos fibre, one wall of which was fitted with the quartz window *Q* directly opposite the spark gap. The opposite wall could be removed for the inspection and removal of the poles, etc., and the diagram shows this wall removed. The length of the spark gap could be adjusted by rotating the threaded brass pole holders *H* and *H′* in their brass sockets *S* and *S′* by means of the ebonite knobs *E* and *E′*. A jet of air could be blown on to the spark in order to cool it by means of the jet *J*. A more constant output from the spark could be obtained in this way. The chimney *C* served to carry off the zinc oxide-laden smoke from the spark. The 6000 V a.c. was applied to the terminals *T* and *T′* and thence to the pole holders *H* and *H′* by wires *W* and *W′* connected to the sockets *S* and *S′*.

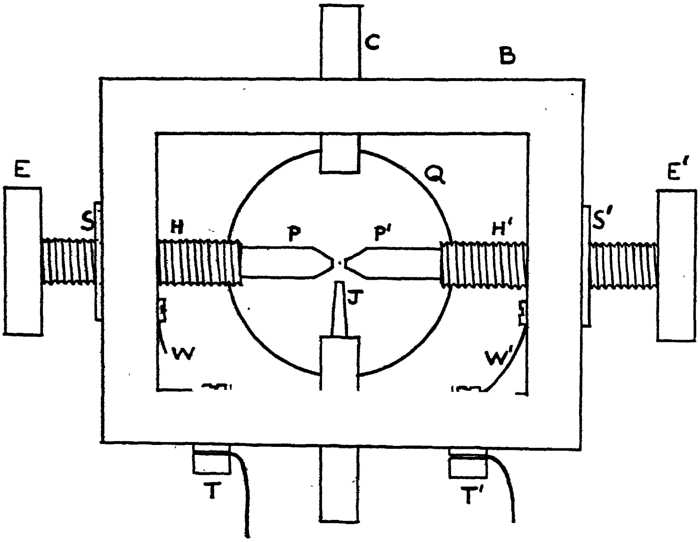


FIGURE 5. Housing for high-voltage zinc spark.

The light from the zinc spark is allowed to pass through a small pressure of ammonia in *A* (figure 4) and then to pass into the vessel *B* where the polymerization takes place. The light partly decomposes the ammonia, and by measuring the amount of non-condensable gas formed the amount of light leaving *A* can be calculated and finally the amount of light entering *B* can be ascertained. Farkas & Harteck (1934) and Wiig (1935), give the figures necessary for these calculations. In this way the light is integrated over a period of time. The polymerization that has taken place is obtained by finding the decrease in pressure* as given by the manometer-barometer arrangement shown. An estimate of the quantum yield can thus be made.

The results are given in table 7.

---

* Before all pressure readings the reaction system was immersed in a water thermostat until equilibrium was attained.

TABLE 7. TEMPERATURE 20° C

run	pressure of vinyl chloride (mm.)	duration of exposure (min.)	number of quanta absorbed by vinyl chloride	number of molecules polymerized	overall quantum yield
1	600	30	$9{\cdot}55 \times 10^{18}$	$2{\cdot}5 \times 10^{19}$	2·6
2	600	10	$1{\cdot}94 \times 10^{18}$	$7{\cdot}1 \times 10^{18}$	3·7
3	600	8	$1{\cdot}0 \times 10^{18}$	$6{\cdot}2 \times 10^{18}$	6·2
4	300	33	$4{\cdot}05 \times 10^{18}$	$2{\cdot}04 \times 10^{19}$	5·0
5	300	30	$6{\cdot}55 \times 10^{18}$	$1{\cdot}57 \times 10^{19}$	2·4
6	100	30	$8{\cdot}18 \times 10^{18}$	$8{\cdot}0 \times 10^{18}$	1·0

Although no quantitative law can be deduced from these results it is apparent that the quantum yield is very small and probably proportional to pressure. The appearance of a cloud of polymer indicates that the chain length is very large, much more than the quantum yield would indicate, which means therefore that the efficiency of the primary photochemical reaction for the starting of chains is very small.

STYRENE

Styrene has a small vapour pressure of about 10 mm. at room temperature. Illumination of 8 mm. for 130 min. at 35° C produced no polymerization and only the slightest trace of non-condensable gas. When 8 mm. were mixed with 50 mm. acetone a rate of about 0·014 mm./min. was obtained corresponding to a chain length of about 3.

METHYL ISOPROPENYL KETONE

Methyl isopropenyl ketone does not polymerize with light of wave-length greater than 2500 A but decomposes to carbon monoxide and a residue. Methyl radicals from acetone, on the other hand, induce it to polymerize. The results at two temperatures are given in table 8. Separate experiments, not recorded, showed that the rate of polymerization is proportional to the 0·6 power of the intensity of the

TABLE 8

temp. °C	run	pressure methyl isopropenyl ketone (mm.)	pressure acetone (mm.)	rate of polymerization (mm./min.)	rate of free radical production (mm./min.)	estimated chain length	CO produced (arbitrary units)	CO expected (arbitrary units)
35	M 8	25	28	0·110	0·0028	40	45	78
	M 15	25	12	0·04	0·0012	33	67	99
	M 17	25	30	0·125	0·003	43	69	101
60	N 6	25	17	0·120	0·004	25	54	120
	N 8	25	28	0·105	0·007	15	49	137
	N 10	26	44	0·195	0·011	18	—	—
	N 11	26	16	0·067	0·004	17	—	—
	N 13	26	31	0·250	0·0074	34	31	75
	N 15	11	31	0·053	0·0074	7	99	179

incident light. Owing to the small saturation pressure of the ketone it is difficult to measure exactly how the rate varies with pressure except to say that it does increase with monomer pressure. Temperature does not seem to have any marked effect on the chain length, and it is of interest to note that the average value of the chain length of the polymer produced at 60° C was about 20, in approximate agreement with the chain-length figures calculated from the data given in table 8. The results in table 8 show that the carbon monoxide production is less than that expected from the decomposition of the acetone and the methyl isopropenyl ketone, once more indicating that acetyl radicals can react with methyl isopropenyl ketone without eliminating carbon monoxide.

The mechanism involved in the production of carbon monoxide from methyl isopropenyl ketone is rather peculiar. Table 9 gives results which when completed show that the rate of carbon monoxide production per mm. of methyl isopropenyl ketone is inversely proportional to the first power of the pressure of the ketone over the range investigated. These observations imply that the following kind of mechanism is involved in the dissociation.

TABLE 9. TEMPERATURE 60° C

run	pressure methyl isopropenyl ketone (mm.)	rate CO production in arbitrary units (per mm. ketone/min.)	reciprocal rate of CO production
0, 2	22	0·090	11·1
0, 4	16	0·106	9·4
0, 6	8	0·132	7·6

Let (K') be the concentration of excited ketone molecules which are deactivated at a rate $k_1(K')(K)$ and $k_2(K')$ respectively by collision and by spontaneous loss of energy. Thus

$$(K') = \frac{I_{\text{abs.}}(K)}{k_1(K)+k_2}.$$

$I_{\text{abs.}}$ is the rate of production of activated molecules per mm. of ketone, the rate of carbon monoxide production is $k_2(K')$ and therefore

$$\frac{d(\text{CO})/dt}{(K)} = \frac{k_2(K')}{(K)} = \frac{k_2 I_{\text{abs.}}}{k_1(K)+k_2},$$

which is in accord with the facts. The methyl isopropenyl ketone molecules must decompose thus

$$\text{CH}_2{=}\text{C}(\text{CH}_3)\,\text{COCH}_3 \rightarrow \text{CH}_2{=}\text{C}(\text{CH}_3)_2 + \text{CO},$$

without producing any appreciable number of free radicals. This explains the fact that there is no appreciable polymerization of the methyl isopropenyl ketone when it is illuminated.

Some experiments were made with hydrogen atoms, but as with butadiene there is no production of polymer of a chain length corresponding to that obtained with methyl radicals. As the results in table 10 show the rate is proportional to the intensity of the light and the decrease in hydrogen pressure is accompanied by a somewhat larger decrease in pressure of the vapour condensable in liquid air. This points not to hydrogenation of the monomer but to the creation of free radicals which unite with each other to give di-ketones.

Table 10

run	temp. °C	pressure methyl isopropenyl ketone (mm.)	pressure hydrogen (mm.)	relative light intensity	observed rate (mm./min.)	change in hydrogen pressure (mm.)	apparent change in methyl isopropenyl pressure (mm.)	B/A chain length
4	20	24	36	I_0	0·21	5·5	7·9	1·4 :1·7
5	20	24	36	$0{\cdot}384I_0$	0·081	4·9	5·2	1·05:1·5
7	55	25	35	I_0	0·17	4·4	9·4	2·2 :2·2
8	80	25	35	I_0	0·07	2·4	4·1	1·7 :1·8

Acrylic nitrile

The absorption of acrylic nitrile in n-hexane is shown in figure 6. Using wavelengths greater than 2500 A and pressures up to 50 mm. at 20° C only small rates of polymerization are obtained, for example, $1{\cdot}5 \times 10^{-2}$ mm./min. The light absorbed at that wave-length, however, is very small, and hence it is not surprising the rate of polymerization is correspondingly small. As a consequence it is possible to study the acetone-sensitized reaction with comparative ease.

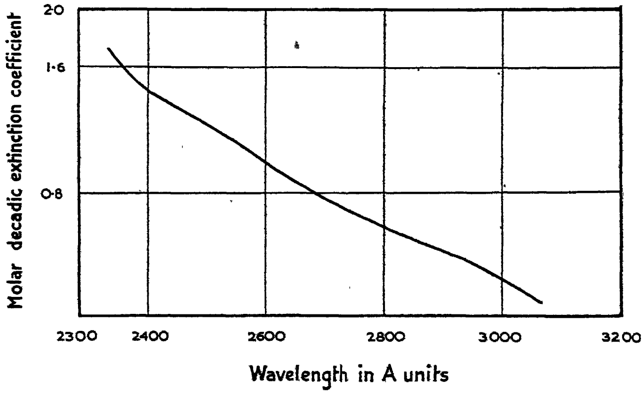


Figure 6. Absorption coefficients of acrylic nitrile in n-hexane.

The acetone sensitized reaction is reproducible and has the characteristics of a free radical mechanism in that the rate has an intensity exponent of nearly 0·5, varies as the square root of the acetone pressure, and as the first power of the acrylic nitrile pressure. Tables 11 and 12 summarize the results.

TABLE 11

run	pressure acrylic nitrile (mm.)	pressure acetone (mm.)	relative intensity	rate observed (mm./min.)	intensity exponent
0, 11	30	32	$0·384I_0$	0·051	0·6
0, 11	29	32	I_0	0·091	
0, 12	30	44	$0·125I_0$	0·033	0·6
0, 12	29	44	I_0	0·113	

TABLE 12

run	pressure acrylic nitrile (mm.)	pressure acetone (mm.)	$\left(\text{pressure acetone}\right)^{\frac{1}{2}}$	rate of acetone sensitized polymerization (mm./min.)	rate of free radical production (mm./min.)	chain length of sensitized polymerization	decomposition observed	decomposition expected
0, 25	31·0	2·5	1·6	0·012	0·00021	57	5·6	6·3
0, 26	31·0	4·8	2·2	0·019	0·00040	48	7·0	8·9
0, 27	31·0	9·6	3·1	0·035	0·00079	44	10·3	12·7
0, 29	31·0	24·0	4·9	0·070	0·00198	35	12·9	19·4
0, 30	31·0	38·0	6·2	0·097	0·00313	31	15·8	24·8
0, 31	31·0	77·0	8·8	0·104	0·00634	16	21·9	46·5
0, 33	9·8	28·0	—	0·042	0·00231	18	41·6	46·0
0, 34	18·3	28·0	—	0·078	0·00231	34	13·1	16·3
0, 37	27·6	28·0	—	0·092	0·00231	40	15·5	21·9

The results are plotted in figures 7 and 8. Since acrylic nitrile itself exhibits a small rate of polymerization this had to be subtracted from the corrected total rate so as to get the acetone-sensitized rate. In figure 7 this rate is plotted against the square root of the acetone pressure and a fair straight line may be drawn through the points, runs 0, 25–0, 31.

Runs 0, 33, 34 and 37 give the variation of the acetone sensitized rate with acrylic nitrile pressure and the results are plotted in figure 8. Here again a linear relationship best represents the results. The estimated chain length as calculated by the rate of free radical production estimated from the rate of carbon monoxide production is given in column 7. As columns 8 and 9 show the carbon monoxide produced with the mixtures was less than that expected if acetone had been alone.

CALCULATION OF $k_p/k_t^{\frac{1}{2}}$

Some estimate of the ratio $k_p/k_t^{\frac{1}{2}}$ can now be made using the above results. It is seen from the equation for the rate of a free radical reaction given at the beginning of this paper that

$$\frac{k_p}{k_t^{\frac{1}{2}}} = \frac{R}{(M)\,(\alpha I_{\text{abs.}})^{\frac{1}{2}}}.$$

A typical rate for a monomer pressure of as near 30 mm. as possible was therefore taken from the above results for each monomer, while the rate of starting of chains

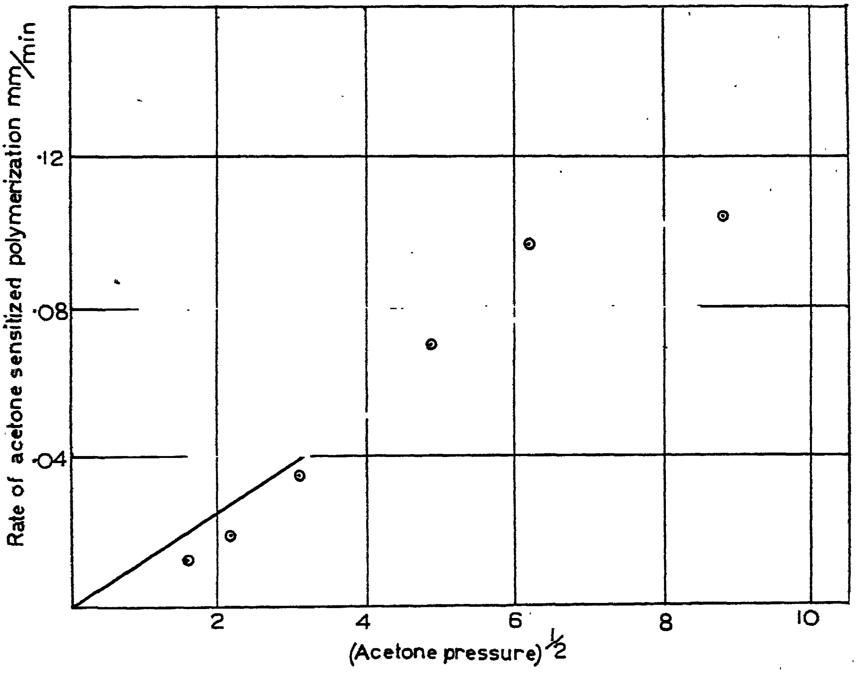


FIGURE 7. Acetone-sensitized polymerization of acrylic nitrile—effect of acetone pressure.

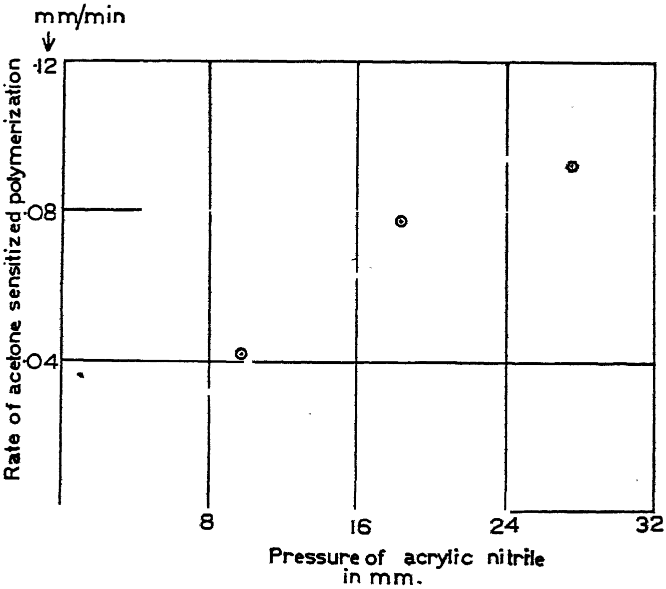


FIGURE 8. Acetone-sensitized polymerization of acrylic nitrile—effect of nitrile pressure.

—$I_{\text{abs.}}$—was given by the rate of free radical production produced with a pressure of as near 30 mm. as possible of acetone in order that a fairly reliable comparison could be made. If the pressure of monomer be expressed in moles/litre and the rates in moles/litre/sec., then values of $k_p/k_t^{\frac{1}{2}}$ for each substance can be obtained. The results are given in table 13.

TABLE 13

temp. °C	substance	pressure monomer (mm.)	concentration (moles/litre) (M)	rate of reaction (moles/litre/sec.) (R)	pressure acetone (mm.)	rate of starting of chains (moles/litre/sec.) ($\alpha I_{\text{abs.}}$)	$(\alpha I_{\text{abs.}})^{\frac{1}{2}}$	$k_p/k_t^{\frac{1}{2}}$
35	butadiene	31	$1{\cdot}61\times10^{-3}$	$1{\cdot}91\times10^{-8}$	21	$1{\cdot}99\times10^{-9}$	$4{\cdot}37\times10^{-5}$	0·27
35	styrene	8	$0{\cdot}42\times10^{-3}$	$1{\cdot}21\times10^{-8}$	50	$4{\cdot}34\times10^{-9}$	$6{\cdot}59\times10^{-5}$	0·44
35	methyl isopropenyl ketone	25	$1{\cdot}30\times10^{-3}$	$10{\cdot}9\times10^{-8}$	30	$2{\cdot}60\times10^{-9}$	$5{\cdot}1\times10^{-5}$	1·63
35	vinyl chloride	31	$1{\cdot}61\times10^{-3}$	$0{\cdot}43\times10^{-8}$	32	$2{\cdot}78\times10^{-9}$	$5{\cdot}27\times10^{-5}$	0·05
20	acrylic nitrile	31	$1{\cdot}70\times10^{-3}$	$6{\cdot}39\times10^{-8}$	24	$1{\cdot}82\times10^{-9}$	$4{\cdot}29\times10^{-5}$	0·88

The values of the ratio $k_p/k_t^{\frac{1}{2}}$ give some indication of the tendency of the molecule to polymerize. It is unfortunate that no results at 35° C were available for acrylic nitrile, but one run done at about 40° C with approximately the same pressure as the run at 20° C decreased the rate by about half. It is estimated therefore that at 35° C $k_p/k_t^{\frac{1}{2}}$ for acrylic nitrile would be of the order 0·50, thus bringing it in line with styrene and butadiene. The relative values of the ratio at 35° C taking vinyl chloride as unity are as follows:

vinyl chloride	butadiene	styrene	acrylic nitrile	methyl isopropenyl ketone
1	5·5	9	(10)	33

It must be remembered that these relative values hold for the free radical polymerization of these substances and they give no indication of the velocity of thermal polymerization.

We should like to thank Bakelite Ltd. for a maintenance grant and the Department of Scientific and Industrial Research for a similar grant to one of us (T.T.J.). We also thank the Chemical Society for a grant for the purchase of silica apparatus. The experimental work described was done in the Colloid Science Laboratory, Cambridge, prior to the outbreak of war and we are very grateful to Professor E. K. Rideal for his encouragement and guidance throughout these researches.

We are indebted to Messrs I.C.I. (Dyestuff Division) Ltd. for kindly providing cylinders of vinyl chloride and of butadiene which materially assisted the carrying out of experiments.

REFERENCES

Jones, T. T. & Melville, H. W. 1940 *Proc. Roy. Soc.* A, **175**, 392.
Farkas, L. & Harteck, P. 1934 *Z. phys. Chem.* B, **25**, 261.
Wiig, E. O. 1935 *J. Amer. Chem. Soc.* **57**, 1559.
Wiig, E. O. 1937 *J. Amer. Chem. Soc.* **59**, 827.

The adsorption of vapours on mercury

I. Non-polar substances

BY C. KEMBALL AND E. K. RIDEAL, F.R.S.

Department of Colloid Science, Cambridge

(*Received* 20 *November* 1945—*Read* 23 *May* 1946)

Reversible results for the adsorption of benzene, toluene and *n*-heptane vapours on mercury have been obtained. The films were found to be gaseous and obeyed the Volmer equation

$$F(A-b)=kT,$$

where F = spreading pressure, A = area per molecule and b = co-area. The possibility that the films might be immobile was considered and the Langmuir equation was applied but found unsatisfactory.

A standard state for the surface phase was defined and the free energy, total energy and entropy of adsorption evaluated. The heat of adsorption was shown to increase with the amount on the surface.

A number of phase changes were found to occur after the completion of monolayer adsorption, the most striking being interpreted as the change over from 'flat' to 'vertical' adsorption of the toluene molecules. Others were thought to be either two-dimensional condensation or adsorption of a second layer.

Much of the published work on the surface tension of mercury in the presence of vapours is qualitative rather than quantitative, and in some cases grave doubts arise as to the purity of the mercury used. The present investigation was undertaken with the intention of throwing further light upon the following points:

(1) The determination of quantitative energies of adsorption of simple molecules on mercury to give the magnitude of the van der Waals or exchange forces involved.

(2) The nature of the monolayer films formed on mercury.

(3) The conditions governing the formation of multilayers or lenses on mercury, which is an ideal medium for examining changes at vapour pressures approaching the saturation value, because of the homogeneity of the surface and the absence of capillaries.

Iredale (1923, 1924 and 1925) described the adsorption of a number of vapours. He found that the adsorption was largely reversible except in the case of methyl iodide. Unfortunately, as pointed out by Burdon (1932), the sessile drop used was not of sufficient size to justify the application of the Worthington (1885) equation to calculate the surface tension, but by chance the errors approximately cancelled.

Thus Iredale's results can only be taken as qualitative data. In all cases the curves for the surface pressure against the vapour pressure were concave to the pressure axis, indicating a departure from perfection of the adsorbed film due to the size of the molecules.

Micheli (1927) obtained a series of curves for the lowering of the surface tension of mercury by the vapour of the hydrocarbons C_5–C_8, again concave to the pressure axis. The effect on the surface tension per additional CH_3 group was approximately three-fold, a factor which has been mentioned in connexion with the adsorption of vacuum grease on mercury (Kemball 1946).

Cassel & Salditt (1931) and Cassel (1931) carried out the most extensive study of the adsorption of vapours on mercury. They worked with the simple alcohols, water, nitromethane, benzene, hexane and *cyclo*hexane. They reported no adsorption of water, and a considerable pressure of the lower alcohols was required in order to produce a noticeable lowering of the surface tension. In fact, they concluded that the polarizability of the molecule was more important for adsorption than dipole moment. However, their initial value for the surface tension of mercury was only 455 dynes/cm. at 50° C, which, coupled with the results to be described shortly for the adsorption of water and the alcohols, points to the correctness of Adam's suggestion that their mercury was contaminated (1941), and further, that the contamination was non-polar in nature.

Bosworth (1932, 1939) adsorbed the lower fatty acids on mercury following the adsorption both by surface-tension measurements and also by dropping mercury through a vapour of known pressure and collecting the acid liberated when the drops coalesced under water. His later work on acetic acid was more extensive, and he determined isotherms over a wide range of temperatures. He gave 8·1 kcal./g.mol. and 6·1 kcal./g.mol. for heats of adsorption at molecular areas of 400 and 25 A^2 respectively. However, these values for acetic acid are not reliable because the association of the vapour was neglected.

EXPERIMENTAL

The method involved measurement of sessile drops and has been described elsewhere (Kemball 1946). The methods of purification and the data for determination of vapour pressure from temperature of the substances adsorbed were as follows:

Benzene. Benzene was shaken with sulphuric acid until free from thiophene; it was then washed with alkali and water, and after drying over sodium, distilled twice. It was kept over sodium until used. Comparison with a specimen from Bureau International Étalons Physico-chimiques, Université de Bruxelles, by ultra-violet spectra, showed that it was over 98 % pure. The physical properties of the comparison specimen were $\eta_D^{20} = 1{\cdot}5013$ and $d_4^{15} = 0{\cdot}88419$. The vapour pressures were calculated from the relation (1)

$$\log_{10} p_{\text{mm.}} = \frac{-0{\cdot}05223a}{T} + b, \qquad (1)$$

using the following constants (*International Critical Tables*):

°C	a	b
−58 to −30 (solid)	42,904	9·556
−30 to − 5 (solid)	34,222	9·846
0 to 42 (liquid)	34,172	7·9622

Toluene. Commercial toluene was shaken with sulphuric acid until there was no further reaction. It was then washed, dried and distilled. The boiling-point was 110·5–111·5° C/760 mm. and $\eta_D^{20} = 1{\cdot}4967$. Vapour pressures were calculated from relation (1) using $a = 39{,}189$ and $b = 8{\cdot}330$ (*International Critical Tables*).

n-Heptane. The heptane used was an hydrogenated Fischer-Tropsch product with $\eta_D^{20} = 1{\cdot}3882$, $(F-C)\times 10^4 = 68$ and $d_4^{20} = 0{\cdot}6845$. When the vapour pressure results of Mundel (1913) (−63 to −40° C) were plotted as $\log p$ against $1/T$, a straight line parallel to the straight line through the results of Young (1898) (0 to 40° C) was obtained. For calculation of the vapour pressure the straight line intermediate between the two sets of results was used.

TABLE 1. *n*-HEPTANE

25° C; surface tension of mercury = 484·2 dynes/cm.

temp. °K of heptane	p (mm.)	F (dynes/cm.)	temp. °K of heptane	p (mm.)	F (dynes/cm.)
202·3	0·0386	5·7	259·0	5·01	42·8
209·3	0·0800	9·4	263·0	6·49	45·7
210·1	0·0875	10·0	265·0	7·43	46·0
215·6	0·152	12·7	273·1	12·22	51·1
216·5	0·164	13·7	273·1	12·22	51·7
226·4	0·406	20·3	274·0	12·66	51·7
226·4	0·406	20·6	279·5	17·74	55·6
229·6	0·528	22·9	279·3	17·56	55·9
236·5	0·978	27·8	284·7	23·55	58·7
240·8	1·33	30·1	288·8	29·07	60·9
240·8	1·33	30·5	293·8	38·46	65·7
248·9	2·47	35·3	296·4	44·00	68·5
250·6	2·77	37·0	297·2	45·86	70·0
254·8	3·74	39·5			
50° C; surface tension of mercury = 479·0 dynes/cm.					
214·2	0·132	4·0	265·2	7·48	31·5
219·6	0·221	6·0	270·8	10·64	34·4
226·1	0·395	8·3	275·1	13·79	37·0
230·1	0·557	10·6	279·8	18·18	39·3
233·4	0·734	12·6	284·8	23·80	42·5
237·3	1·01	14·9	287·3	27·42	43·4
242·1	1·54	16·9	289·1	30·20	44·7
246·3	2·02	19·5	293·0	37·02	46·6
251·5	2·95	22·8	297·5	46·82	48·8
258·7	4·91	27·3			

Isotherms were normally obtained at 25° and at 50° C, but in the case of benzene some measurements were made at 75° C as well. The temperature of the vapour source, vapour pressure and surface pressure are given in tables 1 and 2 for n-heptane and toluene respectively. The results for benzene are given graphically (figure 1), the curves for the other substances being similar in form. All isotherms were completely reversible, and the application of liquid air to the vapour supply brought the surface tension of the mercury back to the original value.

TABLE 2. TOLUENE

25° C; surface tension of mercury = 484·2 dynes/cm.

temp. °K of toluene	p (mm.)	F (dynes/cm.)	temp. °K of toluene	p (mm.)	F (dynes/cm.)
195·9	0·0076	5·0	269·7	5·49	52·4
201·1	0·0141	7·1	273·1	6·82	55·6
206·5	0·0261	9·7	275·8	8·07	58·1
208·4	0·0321	11·4	276·2	8·28	58·4
215·3	0·0662	15·4	278·0	9·23	60·0
221·6	0·123	20·3	280·2	10·55	62·2
223·1	0·142	21·0	281·6	11·47	63·2
226·7	0·199	23·6	282·4	11·95	63·8
230·4	0·278	25·9	283·0	12·47	64·7
231·9	0·317	26·9	285·6	14·50	66·3
231·7	0·315	27·2	284·4	13·54	66·3
237·8	0·526	31·4	286·3	15·11	66·6
240·6	0·661	32·7	288·0	16·64	69·7
241·0	0·684	33·4	290·0	18·64	71·3
244·7	0·918	36·6	293·1	22·14	75·7
245·9	1·01	37·3	293·5	22·61	76·6
247·9	1·18	38·6	294·1	23·38	76·6
250·5	1·44	39·5	295·0	24·56	78·4
254·2	1·89	42·5	295·7	25·51	79·7
256·7	2·26	44·1	295·8	25·64	79·7
258·2	2·52	45·0	296·1	26·10	80·3
258·4	2·55	45·3	297·5	28·10	80·3
261·9	3·26	47·3	297·1	27·69	81·8
265·4	4·13	49·5	297·9	28·76	82·8
269·2	5·31	52·4	297·6	28·24	83·4
50° C; surface tension of mercury = 479·0 dynes/cm.					
208·9	0·0339	1·3	257·2	2·34	21·1
216·1	0·0718	3·0	257·9	2·47	21·8
219·4	0·0998	4·0	264·2	3·81	26·4
223·2	0·144	5·0	266·6	4·48	28·0
222·0	0·187	6·7	270·2	5·66	29·9
226·7	0·219	7·0	272·6	6·61	32·2
231·5	0·321	8·7	277·7	9·08	35·4
234·0	0·381	9·6	282·1	11·37	39·3
242·0	0·742	13·0	285·4	14·34	41·5
244·9	0·934	14·9	289·3	17·93	44·4
247·4	1·14	16·2	293·1	22·14	46·9
249·2	1·30	17·2	298·5	29·63	50·7
251·1	1·50	18·2			

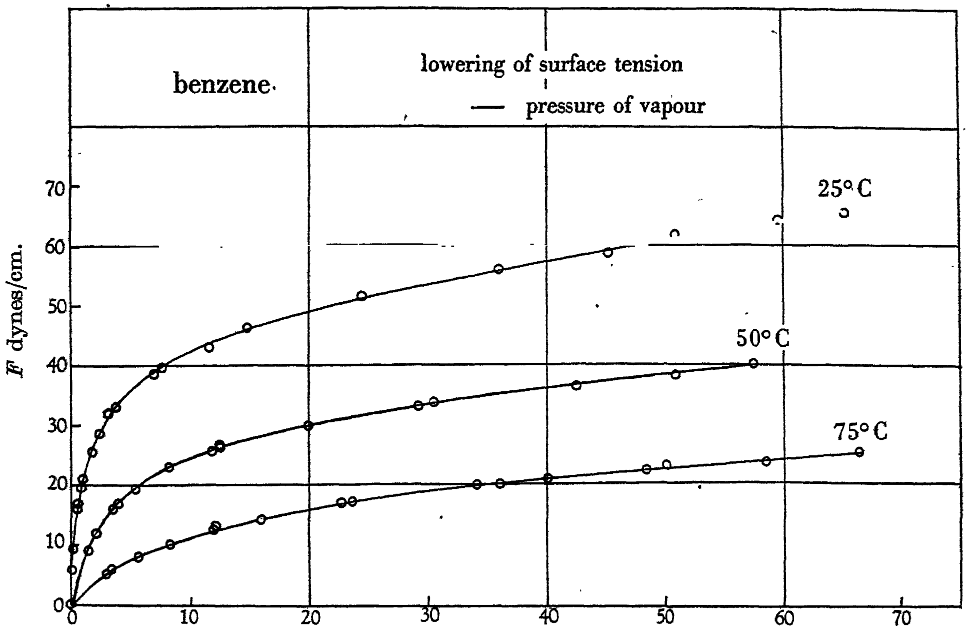


FIGURE 1

INTERPRETATION OF THE EXPERIMENTAL RESULTS

Thermodynamic quantities for the adsorption

In order to obtain the standard thermodynamic quantities it was necessary to consider the surface phase at such concentrations that the molecules could be assumed an ideal gas, and to decide on a standard state for the surface phase. In the past this had been done by considering the surface to have a thickness τ, 5·4 or 6·0 A; it was then possible to convert the two-dimensional surface pressure to a three-dimensional pressure, and to take the standard state as 1 atm. The standard free energy of adsorption is then given by (2),

$$-\Delta G = RT \ln \frac{p_2}{p_1}, \tag{2}$$

where p_1 referred to the bulk phase and p_2 to the surface phase. This could be combined with the Gibbs adsorption equation (3),

$$\Gamma = \frac{1}{RT} \frac{\partial F}{\partial \ln p_1}, \tag{3}$$

where F is the surface pressure and Γ the surface excess. The only assumption involved in this equation is that the bulk phase is ideal, since pressures are used instead of activities. Furthermore, the bulk phase being a vapour, it was possible to ignore the difference between surface excess and surface concentration.

Transforming Γ into suitable units and taking $\tau = 6{\cdot}0\,\text{A}$, (2) and (3) give (4) ($p$ being measured in mm. of Hg),

$$-\Delta G = RT \ln 12{,}500 \frac{\partial F}{\partial p}. \tag{4}$$

This equation is applicable to a region of the adsorption isotherm where the adsorbed film is ideal, in fact the initial value of $\partial F/\partial p$ is required.

The method of defining the standard state by the introduction of an arbitrary surface thickness, which is taken to be the same for all substances, appears at first sight an over-simplification. However, it is equivalent to defining the standard state on the surface as a given area per molecule, which is constant for all molecules at the same temperature, but which depends on temperature. The three-dimensional standard state is usually 1 atm. at any temperature. At $273{\cdot}1^\circ\,\text{K}$ 1 g.mol. occupies 22·4 l. and the corresponding volume per molecule is $36{,}930\,\text{A}^3$, or at any temperature $T^\circ\,\text{K}$ the volume per molecule is given by

$$v = 135{\cdot}2T\,\text{A}^3.$$

Taking the surface thickness as 6 A, the standard surface area per molecule is given by

$$A = 22{\cdot}53T\,\text{A}^2,$$

or, assuming the ideal relation $FA = kT$, the standard surface pressure F is

$$F = 0{\cdot}06084\,\text{dyne/cm.}$$

This last means of defining the standard state is independent of temperature in the same manner as the standard pressure of 1 atm. is independent of temperature.

The experimental curves for F against p were not linear even at the lowest surface pressures measured. This, and the fact that the relative error in F was greatest at small values of F, implied large errors in any graphical method of determining the initial slope. An alternative method was therefore required to obtain the result from the curve as a whole.

The total energy of adsorption could be determined from the free energy at two or more temperatures using the Gibbs-Helmholtz equation (5),

$$\Delta G - \Delta H = T\frac{\partial \Delta G}{\partial T} = -T\Delta S. \tag{5}$$

Although the free energy and the entropy were dependent on the value assumed for τ, the total energy was independent of such an assumption. This can be seen by substituting in (5) from (4) in the form

$$-\Delta G = RT \ln \frac{B}{\tau}\frac{\partial F}{\partial p},$$

where B is a constant, which leads to

$$\Delta H = RT^2 \frac{\partial \ln (\partial F/\partial p)}{\partial T}. \tag{6}$$

The equation of state

The surface monolayers may be either mobile or immobile in the sense that the adsorbed molecules may have two-dimensional freedom or be retained at fixed sites. In the case of activated two-dimensional migration, the apparent mobility is time dependent. Now examine the case on the assumption that the surface layer consists of a mobile film.

The general equation of state for a gaseous surface film is a modification of the Traube equation suggested by Volmer (1925), involving a molecular co-area,

$$F(A-b) = kT. \tag{7}$$

A combination of equation (7) with the Gibbs equation (3) yields

$$\ln p = \ln F + \frac{bF}{kT} + c, \tag{8}$$

containing two constants b and c,

$$p = Fe^{(bF/kT)+c},$$

$$\frac{\partial p}{\partial F} = \left(\frac{bF}{kT}+1\right) e^{(bF/kT)+c},$$

therefore

$$\frac{\partial p}{\partial F_{F\to 0}} = e^{c} \quad \text{or} \quad \frac{\partial F}{\partial p_{F\to 0}} = e^{-c}. \tag{9}$$

The significance of b has been mentioned, but c is also of interest, being some measure of the free energy of adsorption. Considering two substances at the same vapour pressure and having identical values of b, then the difference between F_1 and F_2 for the substances will depend on the difference between c_1 and c_2, the smaller algebraically the value of c, the greater the value of F for a given pressure. Thus the initial slope of the F, p curve is a function of c, equation (9). In order to test equation (8), $\log p/F$ was plotted against F which gave straight lines of slope $b/2{\cdot}303kT$ and intercept c. As can be seen from figures 2–4, reasonable straight lines were obtained over a wide range of surface pressure. For the present only the first region of the isotherm will be considered, and the other regions where changes in the $\log p/F$ against F line occur at higher surface pressures will be discussed later.

The experimental error of the function $\log p/F$ increases at low surface pressure, and therefore all points below about 3 dynes/cm. were ignored. To the remainder of the experimental results, the best straight lines were fitted by the method of least squares. The errors in b and c were calculated similarly. From the values of c, the standard free energies of adsorption were determined using equation (4) in the form (10),

$$\begin{aligned} -\Delta G &= RT \ln 12{,}500 \frac{\partial F}{\partial p_{\text{mm.}}} \\ &= 2{\cdot}303RT \log_{10} 12{,}500 \times 10^{-c} \\ &= 2{\cdot}303RT(4{\cdot}0969 - c). \end{aligned} \tag{10}$$

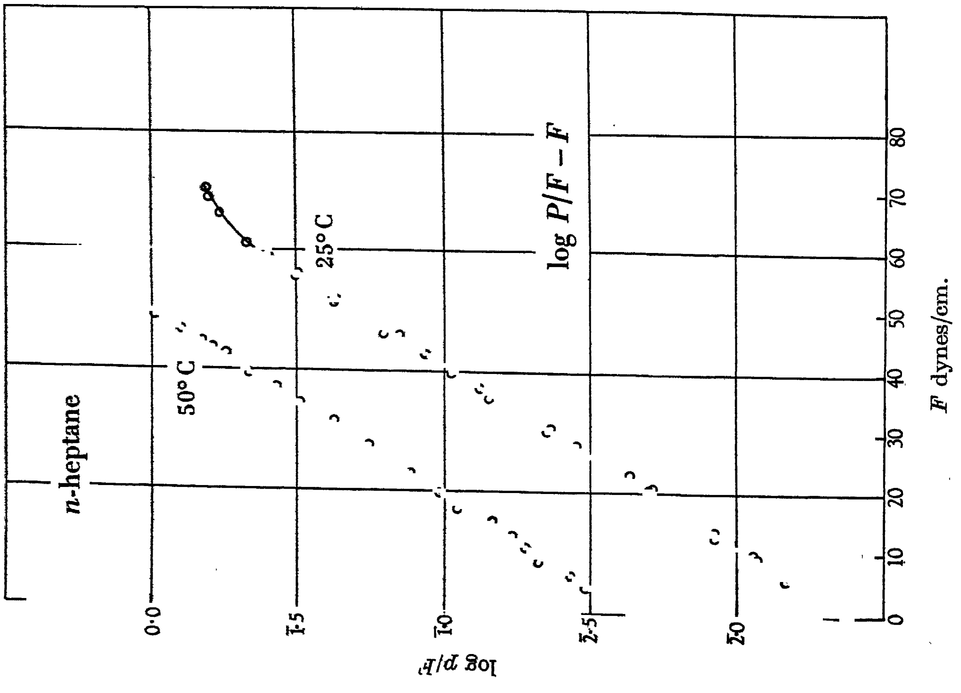


FIGURE 2

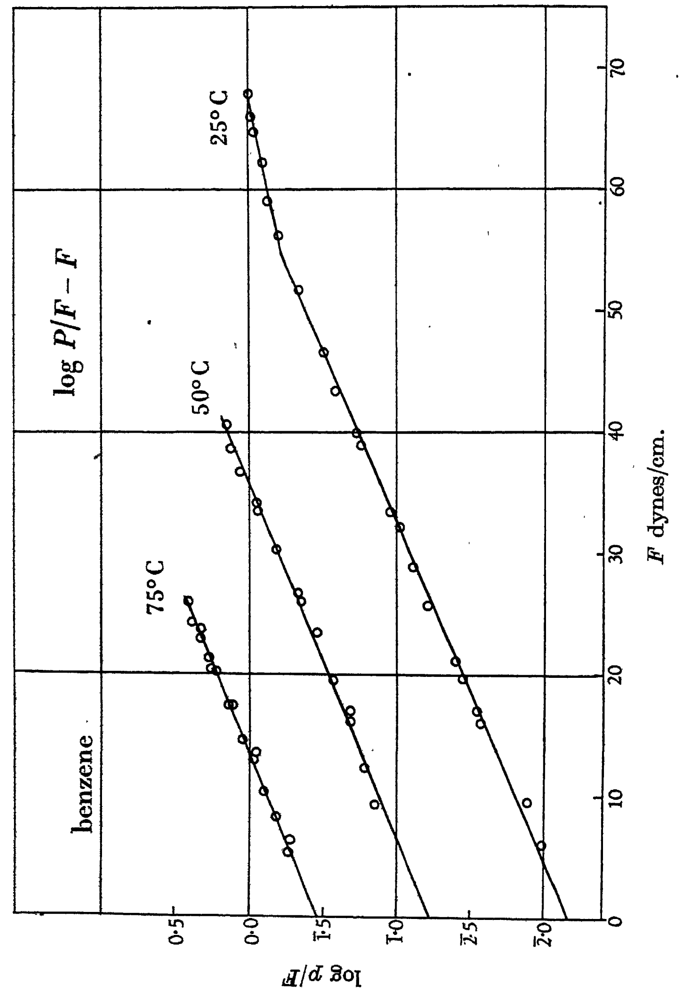


FIGURE 3

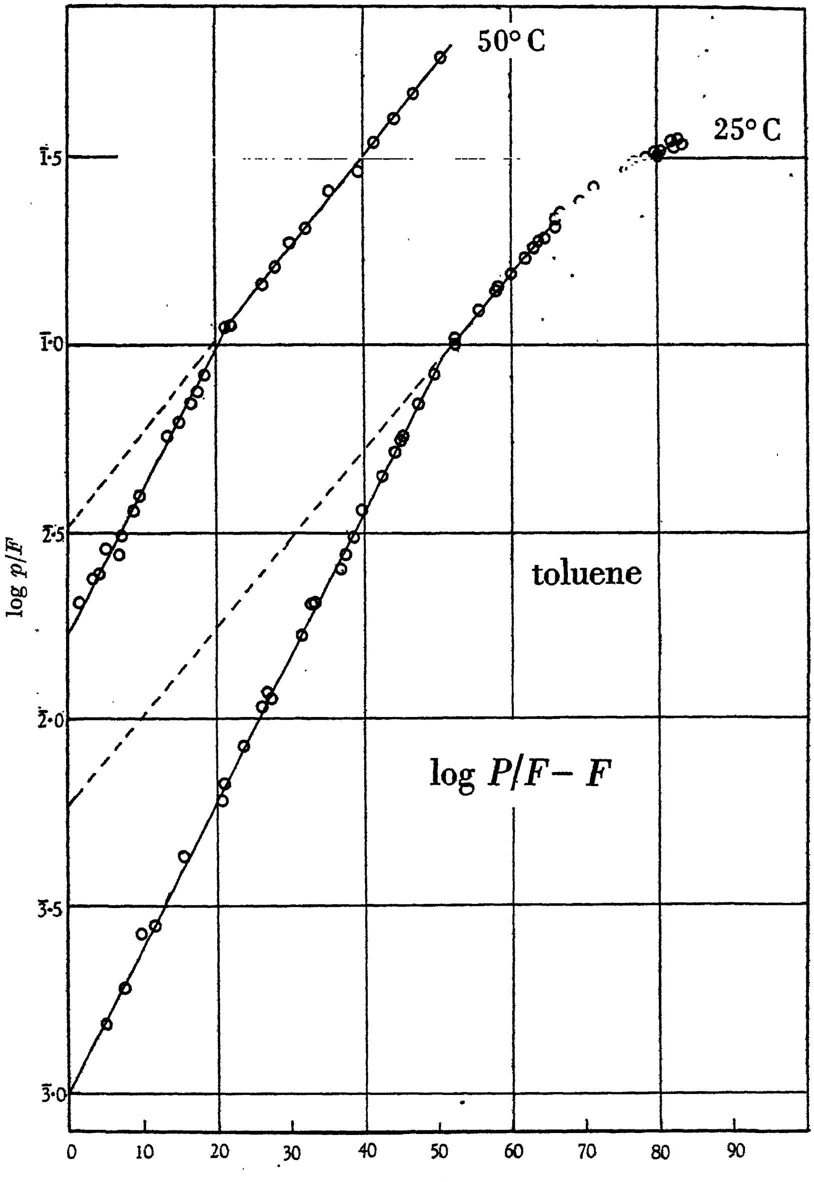


FIGURE 4

The total energy and entropy of adsorption were found from the free energies at two temperatures or more directly from (11), a combination of (6) and (9),

$$\Delta H = 2{\cdot}303RT^2\frac{\partial c}{\partial T}. \tag{11}$$

The values calculated for the free energy, etc., of all the substances are summarized in table 3.

TABLE 3. FREE ENERGIES, TOTAL ENERGIES, ENTROPIES OF ADSORPTION AND CO-AREAS FOR BENZENE, TOLUENE AND *n*-HEPTANE

substance	temp. °C	co-area A²	ΔG cal./g.mol.	ΔH k.cal./g.mol.	ΔS cal./deg.mol.
benzene	25	33·7 ± 0·3	−8513 ± 15	−16·4 ± 0·46	−26·6 ± 1·5
	50	35·0 ± 0·5	−7848 ± 22	−15·7 ± 0·45	−23·9 ± 1·3
	75	37·8 ± 0·5	−7250 ± 11		
toluene	25	36·2 ± 0·2	−9625 ± 8	−21·2 ± 0·30	−39·6 ± 0·9
	50	38·4 ± 0·6	−8634 ± 14		
n-heptane	25	32·1 ± 0·2	−8817 ± 8	−13·4 ± 0·22	−15·3 ± 0·7
	50	33·3 ± 0·2	−8434 ± 9		

Again examine the case on the assumption that the surface film is immobile, the adsorbed molecules being held on sites. In the simplest case where no interaction between neighbouring particles is assumed the Gibbs equation can be combined with that of Langmuir, which relates the surface concentration to the pressure of the bulk phase.

The empirical equation (12) given by Szyszkowski (1908) has been shown by Rideal (1926) to imply the Langmuir relation (13),

$$F = \gamma_0 - \gamma = B\gamma_0 \log_{10}\left(\frac{p}{A}+1\right), \tag{12}$$

$$\Gamma = \frac{B\gamma_0}{2{\cdot}303RT}\frac{p}{p+A}. \tag{13}$$

The method used to fit equation (12) was due to Szyszkowski (1908), i.e. to take fixed intervals, d, of $1-\gamma/\gamma_0$, and to find the values of p_1, p_2, etc., corresponding to these values of $1-\gamma/\gamma_0$. The value of R given by

$$R = \frac{p_n - p_{n-1}}{p_{n-1} - p_{n-2}}$$

should be constant, and the constants A and B in (12) are given by

$$A = \frac{1}{n}\sum_{n=n}^{n=1}\frac{p_n}{R^n-1} \quad \text{and} \quad B = \frac{d}{\log R}.$$

The maximum adsorption, Γ_{lim}, corresponding to a completed monolayer, is

$$\Gamma_{\text{lim}} = \frac{B\gamma_0}{2{\cdot}303RT},$$

and the area per molecule is $1/\Gamma_{\text{lim}}$.

The results for *n*-heptane at both 25° and 50° C were subjected to this treatment, and the values of R are shown in tables 4 and 5. Heptane was chosen because of the existence of the monolayer over a comparatively large range of surface pressures. It can be seen that R decreases throughout the adsorption, particularly in the initial stages where the calculated values of log p differ most markedly from the observed. This means either that the equation was not valid or else that the packing of the molecules was constantly changing, i.e. the co-area was becoming less as the film was compressed. With heptane this explanation is quite possible because at low surface pressures one would have expected the molecules to lie flat on the surface and at high most of each molecule would have been forced off the surface. The mean limiting areas per molecule are 42·7 A^2 at 25° C and 47·0 A^2 at 50° C, which are larger than the corresponding areas determined by the Volmer equation. The difference is connected with the basic assumptions. The Langmuir equation has been shown to apply when the molecules are adsorbed on to definite sites and do not interact with each other (Fowler & Guggenheim 1939), whilst the Volmer equation depends essentially on the mobility of the molecules; thus the resistance to adsorption for the last stages of the monolayer is much greater than the resistance experienced in the case of Langmuir adsorption. Molecules in motion effectively cover a larger area of the surface. As a result over a comparatively large range of surface concentration

TABLE 4. ADSORPTION OF HEPTANE AT 25° C

γ/γ_0	F (dynes/cm.)	log p	p (mm.)	$p_n - p_{n-1}$	$R = \frac{p_n - p_{n-1}}{p_{n-1} - p_{n-2}}$	log p (calc.)
1·00	0·00	—	0·0000	—	—	—
0·99	4·84	$\bar{2}$·480	0·0302	0·0302	—	$\bar{2}$·5514
0·98	9·68	$\bar{2}$·936	0·0863	0·0561	1·854	$\bar{2}$·9765
0·97	14·53	$\bar{1}$·267	0·1849	0·0986	1·753	$\bar{1}$·2856
0·96	19·37	$\bar{1}$·547	0·3524	0·1675	1·700	$\bar{1}$·5519
0·95	24·21	$\bar{1}$·807	0·6412	0·2888	1·722	$\bar{1}$·7980
0·94	29·05	0·053	1·130	0·4888	1·691	0·0330
0·93	33·89	0·297	1·982	0·852	1·739	0·2621
0·92	38·74	0·529	3·381	1·399	1·642	0·4878
0·91	43·58	0·749	5·610	2·229	1·592	0·7114
0·90	48·42	0·954	8·995	3·385	1·518	0·9339
0·89	53·26	1·154	14·26	5·265	1·553	1·1556
0·88	58·10	1·347	22·23	7·97	1·512	1·3767

Mean $R = 1·662$

TABLE 5. ADSORPTION OF HEPTANE AT 50° C

γ/γ_0	F (dynes/cm.)	log p	p (mm.)	$p_n - p_{n-1}$	$R = \frac{p_n - p_{n-1}}{p_{n-1} - p_{n-2}}$	log p (calc.)
1·00	0·00	—	0·0000	—	—	—
0·99	4·79	$\bar{1}$·214	0·1637	0·1637	—	$\bar{1}$·2548
0·98	9·58	$\bar{1}$·663	0·4603	0·2966	1·813	$\bar{1}$·6799
0·97	14·37	$\bar{1}$·994	0·9864	0·5261	1·771	$\bar{1}$·9870
0·96	19·16	0·271	1·866	0·8796	1·668	0·2553
0·95	23·95	0·523	3·334	1·468	1·669	0·5014
0·94	28·74	0·757	5·715	2·381	1·621	0·7364
0·93	33·53	0·978	9·506	3·791	1·586	0·9655
0·92	38·32	1·196	15·70	6·194	1·628	1·1912
0·91	43·11	1·408	25·59	9·89	1·592	1·4148
0·90	47·90	1·618	41·50	15·91	1·608	1·6373

a Volmer expression with co-area 33·3 A^2 gives the same result as a Langmuir expression with co-area 47·0 A^2, i.e. the results for heptane at 50° C. To illustrate the different ways in which the adsorptions approach saturation, the calculated adsorptions for heptane obeying the Szyszkowski equation (14), the equation (15) which was derived from the Volmer equation, and lastly the experimental values calculated by the Gibbs equation are shown in figure 5. The experimental results

$$F = 0{\cdot}04533 \times 479{\cdot}0 \log\left(\frac{p}{0{\cdot}2716} + 1\right), \tag{14}$$

$$\log p/F = 0{\cdot}03259F - 1{\cdot}6248, \tag{15}$$

clearly favour the Volmer adsorption isotherm.

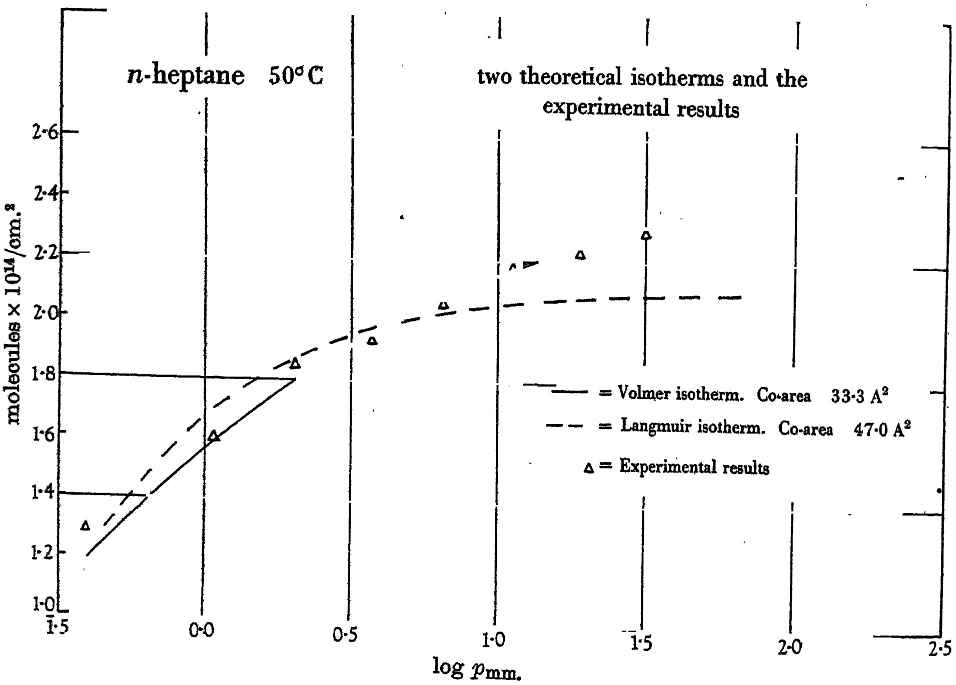


FIGURE 5

The heat of adsorption was obtained from the Szyszkowski equation. Differentiating (12)

$$\frac{\partial F}{\partial p_{p\to 0}} = \frac{B\gamma_0}{A}, \tag{16}$$

and substituting in (6) from (16),

$$\Delta H = RT^2 \frac{\partial \log_e (B\gamma_0/A)}{\partial T}. \tag{17}$$

Equation (17) gives the heat of adsorption of heptane as 12·5 kcal./g.mol. This is lower than the Volmer value of 13·4 kcal./g.mol. because the Langmuir isotherm is curved when plotted as $\log p/F - F$, and because of the greater range to which the Langmuir equation was fitted in the case of adsorption at 25° C. These two factors reduce the difference between the values of $\log p/F$ at $F = 0$ for the two temperatures.

It is clear that in all the cases where straight lines are obtained by plotting $\log p/F$ against F, a Langmuir equation could be fitted, and the constant R would decrease as the adsorption proceeds. The heats of adsorption calculated from such Langmuir equations would not differ greatly from those obtained by the use of the Volmer equation, but the co-areas would be decidedly greater.

The force area curves

The methods used to interpret the results in the preceding sections have all depended on the Gibbs equation and some other relation between the surface pressure, surface concentration or pressure of the bulk phase. However, by the use of the Gibbs equation alone it is possible to construct the force-area curves for the adsorbed substances, and although these do not give much further information about the region of monolayer adsorption, they furnish valuable information about the phase changes and the formation of second layers, which occur at higher surface pressures.

The surface pressure is a direct observation, and the area per molecule is calculated from the curves for F against $\log p$, using the Gibbs equation (3). An example of the F against $\log p$ curves is given in figure 6. The slope changes rapidly at first and then more slowly as the monolayer is completed, and further rapid changes occur at higher pressures which will be considered below.

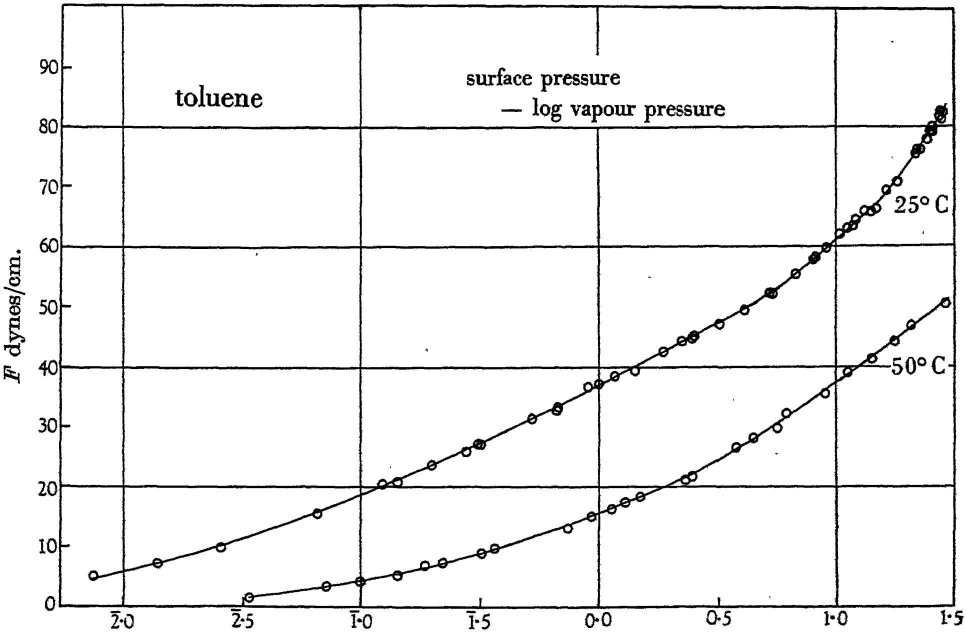


FIGURE 6

The FA–F curves for heptane and toluene are given in figures 7 and 8. The first portion of the curve is in all cases a straight line intercepting the FA axis at kT as required by the Volmer equation. The increase of the co-area, i.e. the slope of this straight portion of the curve, with temperature is also obvious. The main deduction from these curves is that the films are gaseous, because, although there may be a tendency for the value of FA to dip below 400, as F is decreased, it certainly does not tend to zero.

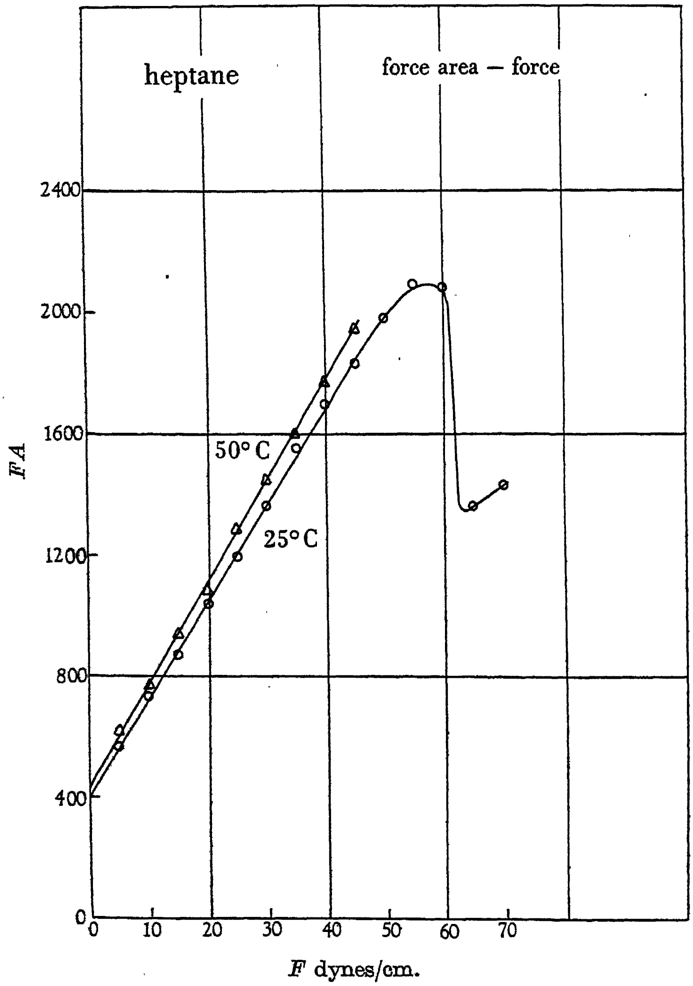


FIGURE 7

On the variation of the heat of adsorption

The investigation of the change in ΔH as the surface becomes covered indicates the magnitude of the interactions between the molecules in the adsorbed film. The calculation is carried out with the experimental results for heptane, the force-area relations (18) and (19) are assumed to hold, the values of the constants being those determined by the Volmer method. The values of $\log p$ for any particular value of F,

$$F(A-32{\cdot}07) = 408{\cdot}6 \text{ at } 25^\circ \text{ C}, \tag{18}$$

$$F(A-33{\cdot}26) = 442{\cdot}9 \text{ at } 50^\circ \text{ C}, \tag{19}$$

are determined from the smoothed curves of F against $\log p$.

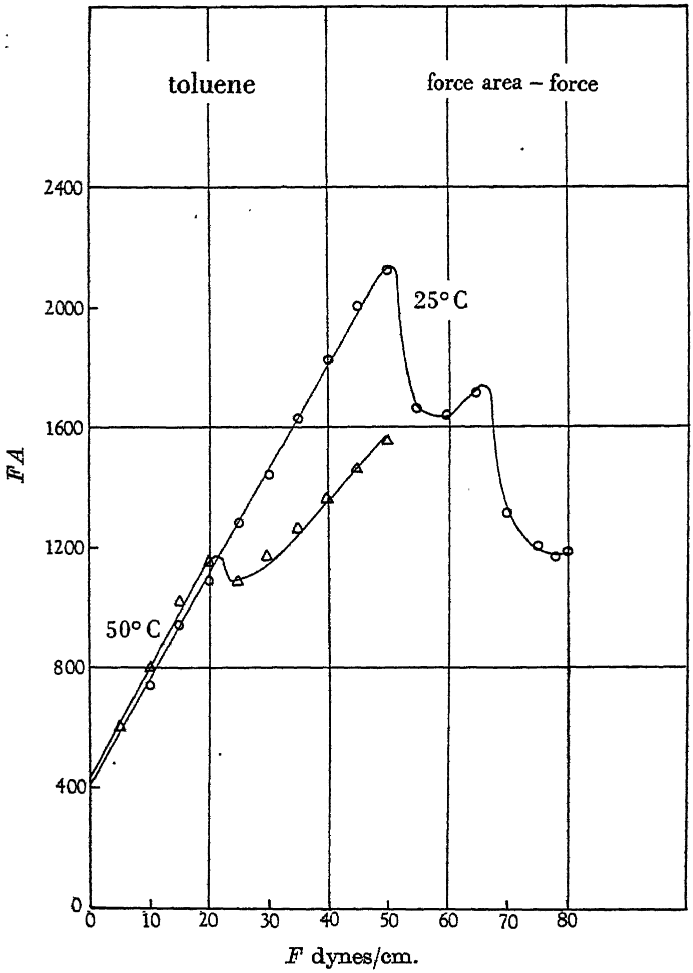


FIGURE 8

The first method of examining the change in ΔH is to calculate the isosteric heat of adsorption for fixed amounts on the surface using the van't Hoff isochore in the form (20)

$$\left(\frac{\partial \ln p}{\partial T}\right)_A = -\frac{q}{RT^2}, \tag{20}$$

where A refers to the area per molecule. The results obtained in this manner are given in table 6. The energy of adsorption increases slowly at first and then more rapidly towards the completion of the monolayer. This indicates that energy is liberated by the approach of two heptane molecules to their equilibrium distance as expected.

TABLE 6. ISOSTERIC HEAT OF ADSORPTION OF HEPTANE

surface concentration 10^{-13} mol./cm.^{2}	10	15	20	23·24
area per mol. A^2	100	66·7	50	43
heat of adsorption, kcal./g.mol.	14·37	14·79	15·86	18·42

The second method of calculating the variation in the heat of adsorption is to refer the change in $\log p$ with temperature to the fraction of the surface covered. The surface is assumed to be covered when there is one molecule per 32·07 A^2 at 25° C and per 33·26 A^2 at 50° C. The results for this method of calculation are given in table 7. In this case, allowance having been made for the change in co-area with temperature, the heat of adsorption is constant. This means that the increase in co-area with temperature results from the liberation of energy when two molecules approach to their equilibrium distance. The two sets of results in tables 6 and 7 confirm the conclusions reached by Wilkins (1938), who recommended from statistical mechanical arguments that the heat of adsorption should be calculated with respect to the fraction of the surface covered and not with respect to the amount on the surface. The difference between the result of 13·77 cal./g.mol., the mean of the results in table 7 and the value 13·4 cal./g.mol., obtained earlier by the use of the Volmer equation, is due to the fact that the first is a heat at constant volume, and the second at constant pressure.

TABLE 7. VARIATION IN THE HEAT OF ADSORPTION OF HEPTANE WITH FRACTION OF SURFACE COVERED

fraction covered	0·3	0·4	0·5	0·6	0·7	0·8
heat of adsorption, kcal./g.mol.	13·78	13·73	13·80	13·73	13·73	13·83

DISCUSSION OF THE RESULTS FOR THE MONOLAYER REGION

Calculations both by the Volmer equation and the Gibbs equation show that the first region indicated by the $\log p/F$ curves corresponded to monolayer adsorption. The three substances were found to obey the Volmer equation (7), and the co-area was slightly dependent on temperature. The Langmuir equation was applied to the results for heptane and did not give such good agreement. The monolayers appeared to be gaseous, as the value of FA did not approach zero as F was decreased. The co-areas determined compared favourably with areas calculated from the molecular volume of the molecules in the liquid state (cf. table 8). The experimental co-areas for benzene and toluene were greater than the value from the liquid volume because they were flat on the surface. Palmer & Clark (1935) gave 36 A^2 for the low-pressure area of benzene on vitreous silica, and Stewart (1929) obtained 4·70 A for the thickness of benzene in the liquid state by X-ray diffraction, hence 31·4 A^2 for the area, in agreement both with 34·4 A^2 for benzene and 37·3 A^2 for toluene. If the *n*-heptane molecules had been fully extended on the surface, they would have required about 50 A^2, and the observed value of 32·7 A^2 indicated that the molecules were partially curled up, having four or five of the carbon atoms on the surface. It is possible with models to make a heptane structure which has the same area as a benzene ring and almost five of the carbon atoms in one plane. The increase in co-area with temperature has been discussed above and shown to imply liberation of energy by the approach of two molecules. The co-areas determined by the Volmer method were assumed to represent the actual area of the molecules, not twice the area as expected from the simple theory of molecules in a dilute gas (cf. the three-dimensional van der Waals constant *b* is four times the volume of the molecules). The simple treatment must fail as soon as the concentration of the gas is sufficiently great to get a large number of multiple collisions and must be inapplicable as soon as the surface is half covered or else infinite surface pressure would be required to satisfy (7). The reversibility of the adsorptions and the magnitude of the heats indicated that only van der Waals's interaction was involved. The value of 13·4 kcal./g.mol. for *n*-heptane indicates the comparatively large quantity of energy that can be liberated from the adsorption of non-polar molecules. The adsorption of a benzene ring leads to an even higher heat of adsorption. This effect is not due to the additional polarizability of the benzene nucleus but to the fact that the molecule is planar and can approach more closely to the surface before the short-range repulsive forces take charge. The free energies of adsorption showed considerably less variation than the total energies.

TABLE 8. CO-AREAS AND AREAS FROM THE LIQUID STATE

substance	liquid volumes A^3	$(\text{volume})^{\frac{2}{3}}$ A^2	co-area A^2
benzene	146·6	27·8	34·4
n-heptane	241·6	38·8	32·7
toluene	175·6	31·4	37·3

The extensive work by Cassel & Salditt (1931) has been quoted and the possibility of contamination of their mercury mentioned. With non-polar substances their curves for F against p were concave to the pressure axis; with water no adsorption occurred and with nitromethane a large pressure was required before any lowering of surface tension was detected, and then a fairly rapid lowering occurred with further increase of pressure. The alcohols showed serial changes, methyl requiring a high pressure to start affecting the surface tension, ethyl a somewhat lower pressure and so on until butyl gave an F against p curve entirely concave to the pressure axis. In all cases once the critical pressure had been reached further increase of pressure gave marked adsorption. Their mercury must have been covered with a non-polar contamination with which non-polar compounds were completely miscible, and in which the alcohols could be adsorbed once a critical pressure was reached which was lower the greater the amount of hydrocarbon in the alcohol molecule. In preliminary experiments with water and the lower alcohols no initial flat portion of the $F-p$ curve has been found.

The significance of the entropy values have not been discussed because it is intended to study them in conjunction with theoretical calculations to be described in the following paper.

PHASE CHANGES AND SECONDARY LAYERS

Monolayer adsorption has been a recognized phenomenon for some years, but whether the phenomenon of condensation involves the passing of a monolayer by stages into a 'multilayer' at relatively low values of p/p_0, or whether the monolayer persists until very near the saturation pressure, and then changes abruptly into three-dimensional condensation, has not been fully explored. One difficulty has been the porous nature of many adsorbing materials, and a second, affecting the onset of multilayer formation, the inhomogeneity of the adsorbing surfaces. Liquid mercury suffers from neither of these defects, and particular interest arises as to the interpretation of the changes observed in the isotherms at higher values of p/p_0. The work of Palmer & Clark (1935) and Palmer (1937, 1938) on vitreous silica and of Orr (1939) on alkali halides indicated the adsorption of three to four molecular layers for high values of p/p_0. Conclusions reached about adsorption on mercury cannot be applied without modification to solid surfaces, but where multilayers form on mercury they are more likely to be formed on a solid surface of the same adsorbing power because of the more permanent surface of a solid surface.

In the case of n-heptane and mercury the monolayer region persists until a fairly high surface pressure and then further adsorption occurs. This might be due either to two-dimensional condensation, to the adsorption of a second layer, or to the unrolling of the heptane molecules and adsorption end-on to the surface. The positions of the last two or three points of the $FA-F$ curve (figure 7) depend on the way the $F-\log p$ is drawn and cannot be regarded as accurate. The FA value is about 1400 at 70 dynes/cm., whereas if the molecules had unrolled to a co-area of 20 A, the expected value would have been 1800. If a second layer of molecules, obeying a Volmer equation, were adsorbed on the first layer, the two layers might be assumed to exert separate surface pressures. If x be the fraction of the adsorbed material in the first layer and A the area per molecule from the Gibbs equation, then

$$F_1\left(\frac{A}{x}-b\right) = kT, \tag{21}$$

$$F_2\left(\frac{A}{1-x}-b\right) = kT, \tag{22}$$

where F_1 and F_2 refer to the surface pressures of the two layers. Since

$$F_1 + F_2 = F$$

we obtain
$$F(A - \tfrac{1}{2}b) = kT - (F_1 - F_2)\, b(\tfrac{1}{2} - x). \tag{23}$$

In the initial stages of the adsorption there is no second layer and the ordinary Volmer equation obeyed, i.e. (23), with $x = 1$. Once the first gaseous layer is sufficiently closely packed a second layer may be formed. It is assumed that this cannot occur until the fraction of the surface covered is fairly high or else molecules in the second layer would simply fall into the first layer. Once started, the second layer begins to fill up rapidly, i.e. x tends to $\frac{1}{2}$ fairly rapidly, and the last term of (23) becomes smaller with respect to kT. The relation (23) would then reduce approximately to the form (24),

$$F(A - \tfrac{1}{2}b) = kT. \tag{24}$$

On this view the expected value of FA for heptane at $F = 70$ dynes/cm. would be 1500.

If the heptane had condensed to a two-dimensional liquid or a vapour-expanded film the FA value would have risen from zero at low-surface pressures and followed a line parallel to the FA for a gaseous film. An example of this for the fatty acids on water was given by Adam (1941). If the heptane molecules unrolled to form a vapour-expanded film with a co-area of 20 A^2, the FA value at $F = 70$ dynes would be about 1400. Thus the explanation of the phenomenon observed with heptane might be either the formation of a second layer or two-dimensional condensation in the first layer.

The results with toluene appeared to indicate a different type of phase change from those found with n-heptane. The equations

$$F(A - 22{\cdot}0) = kT \text{ at } 25^\circ\text{ C} \quad \text{and} \quad F(A - 24{\cdot}9) = kT \text{ at } 50^\circ\text{ C}$$

could be fitted to the second region of the adsorption. The thermodynamic quantities for this second phase (assumed stable down to the standard state) were calculated as for the monolayer region. The free energy of adsorption was -8401 ± 20 cal./g.mol., the heat $-13{\cdot}1 \pm 0{\cdot}5$ kcal./g.mol. and the entropy $-15{\cdot}2 \pm 1{\cdot}6$ entropy units. It is probable that this phase represented the adsorption end-on to the surface. The co-area of 23·5 A^2 agrees with the cross-sectional area of 24 A^2 given by Adam (1941) for an aromatic amine or phenol on a water surface. The energy values also confirm this deduction, and it is possible to apply the two-dimensional Clausius-Clapeyron equation

$$\frac{\partial \ln F}{\partial T} = \frac{-q}{RT^2}$$

obtaining $q = 6{\cdot}7$ kcal. This differs from the 8·1 kcal. found by the methods outlined above because the phases are by no means perfect at the concentrations of the phase change and the Clausius-Clapeyron assumes perfect behaviour. The change which occurs at about 53 dynes/cm. in the isotherm for benzene at 25° C is believed to correspond to the toluene phenomenon.

The further change which occurs in the adsorption of toluene between 65 and 70 dynes/cm. must mean the formation of a second layer. It cannot correspond to two-dimensional condensation because the area per molecule at 80 dynes/cm. is only 15 A^2. However, two layers of co-area 22 A^2 would give $FA = 1300$, agreeing within the experimental error with the 1200 observed. The $\log p/F - F$ curves and the $FA - F$ curves afford two different methods of examining all these changes. In the $\log p/F - F$ curves, the changes are made as sharp as possible by drawing straight lines and the $FA - F$ are the result of smoothed $F - \log p$ curves. Despite experimental error and smoothing the discontinuities remain fairly sharp, a factor presumably being the homogeneous nature of the surface.

Some work by Cassel (1932) on the adsorption of carbon tetrachloride on mercury has attracted considerable attention. He claimed that his results indicated 3·6 molecular layers as the saturation vapour pressure was approached, and, moreover, that the first molecular layer was completed at an extremely low value of p/p_0. However, the general form of his $F - \log p$ curve was similar to those in figure 6, where the monolayer was not complete until a high p/p_0. Some calculations were made from his results. At the point where he claimed the adsorption of about $1\frac{1}{4}$ layers (with 31·8 A^2 as the co-area), i.e. 26 A per mol., calculation gave 137 A^2 per mol. Where he claimed 3·6 layers, i.e. area 9·5 A^2 per mol., calculation gave 50 A^2 per mol. A possible explanation appeared when the ratio of the two sets of results was found to be close to 5·3, which is the square of 2·303, the factor for conversion from natural to logarithms to the base 10. It would appear, therefore, that Cassel's results far from indicating multilayer adsorption do not even represent the completion of a monolayer. However, preliminary work has shown that the adsorption of carbon tetrachloride is complicated by chemisorption and irreversibility, a fact not detected by the earlier experiments using the drop-weight method of measuring surface tension.

The work described in the present paper was financed by the Ministry of Aircraft Production, and we are grateful for the permission received to publish.

REFERENCES

Adam 1941 *The physics and chemistry of surfaces.* Oxford Univ. Press.
Bosworth 1932 *Trans. Faraday Soc.* 28, 903.
Bosworth 1939 *Trans. Faraday Soc.* 35, 1349.
Burdon 1932 *Trans. Faraday Soc.* 28, 866.
Cassel 1931 *Z. Elektrochem.* 37, 642.
Cassel 1932 *Trans. Faraday Soc.* 28, 177.
Cassel & Salditt 1931 *Z. phys. Chem.* A, 155, 321.
Fowler & Guggenheim 1939 *Statistical thermodynamics.* Cambridge Univ. Press.
Iredale 1923 *Phil. Mag.* 45, 1088.
Iredale 1924 *Phil. Mag.* 48, 177.

Iredale 1925 *Phil. Mag.* **49**, 603.
Kemball 1946 *Trans. Faraday Soc.* (in the Press).
Micheli 1927 *Phil. Mag.* **3**, 895.
Mundel 1913 *Z. phys. Chem.* **85**, 452.
Orr 1939 *Proc. Roy. Soc.* A, **173**, 349.
Palmer 1937 *Proc. Roy. Soc.* A, **160**, 254.
Palmer 1938 *Proc. Roy. Soc.* A, **168**, 190.
Palmer & Clark 1935 *Proc. Roy. Soc.* A, **149**, 360.
Rideal 1926 *Surface chemistry*, 1st ed. p. 42. Cambridge Univ. Press.
Stewart 1929 *Phys. Rev.* **33**, 889.
Szyszkowski 1908 *Z. phys. Chem.* **64**, 385.
Volmer 1925 *Z. phys. Chem.* **115**, 253.
Wilkins 1938 *Proc. Roy. Soc.* A, **164**, 496.
Worthington 1885 *Phil. Mag.* **20**, 51.
Young 1898 *J. Chem. Soc.* **73**, 675.

The adsorption of vapours on mercury
II. The entropy and heat of adsorption of non-polar substances

By C. Kemball, *Department of Colloid Science, Cambridge*

(*Communicated by E. K. Rideal, F.R.S.—Received* 20 *November* 1945—*Read* 23 *May* 1946)

An attempt has been made to interpret the entropy of adsorption of benzene, toluene and *n*-heptane on mercury. These investigations have indicated with fair accuracy the amount of translational and rotational freedom possessed by the substances on the surface of mercury. The most striking result was obtained with benzene, where by denying all rotation except in the plane of the ring and denying the third degree of translational freedom, the calculated entropy of adsorption agreed closely with the experimental value. The surface mobility of toluene was found to be considerably hindered, and the entropy of adsorption of *n*-heptane confirmed the view that the molecules were partially rolled up.

An attempt has been made to derive the theoretical heat of adsorption of benzene from various relations for the attractive and repulsive potentials for the van der Waals forces near a metal surface. The values obtained were of the same magnitude as the experimental value. The calculations gave some evidence of the ranges of intermolecular distance over which the different equations for the attractive potential were accurate.

In the preceding paper experimental values for the entropy and heat of adsorption on mercury of benzene, toluene and *n*-heptane were obtained. The present paper describes an attempt to calculate the entropy of adsorption in the above cases and the heat of adsorption of benzene. The entropy can give information about the nature of the adsorbed film and the freedom experienced by the molecules on the surface of mercury, whilst the heats of adsorption can be compared with the expected values due to the calculated van der Waals forces.

THE ENTROPY OF ADSORPTION

The exact calculation of the entropy of substances adsorbed on mercury is not possible because the nature of the force field at the surface is not known. The method used was to consider the molecules confined to a plane at the standard concentration of 1 mol./$22\cdot53T$ A^2 (the standard state defined in the preceding paper). These are assumed to have two degrees of translational freedom in the plane, and the motion perpendicular to the plane is replaced by a vibration. The entropy due to this new vibration will be neglected as a first approximation. Having thus denied the third mode of translation, the various modes of rotation will be denied in turn and the entropy calculated at each stage, and compared with the experimental entropy for the substance adsorbed on mercury under ideal conditions, i.e. the entropy for adsorption to the standard state of 1 mol./$22\cdot53T$ A^2. The conditions which appear to give the same entropy of adsorption for the molecules in the theoretical plane as the experimental value may then be likened to the conditions on the mercury surface. The main criticism of the method is that the denial of the rotational degrees of freedom one by one is rather an ideal case, and that in practice the force field of the mercury is more likely to interfere to some extent with all the rotational possibilities. This criticism will be considered in the individual cases.

The translational entropy in three dimensions is given by the Sackur-Tetrode equation,

$$S_{\text{trans.}} = R\ln(M^{\frac{3}{2}}T^{\frac{5}{2}}) - 2\cdot30, \tag{1}$$

where M is the molecular weight and the entropy is given in cal./degree/g.mol. for the pressure of 1 atm.

In two dimensions the partition function for a single molecule is

$$f_A = \frac{2\pi mkT}{h^2}a,$$

where m is the weight of a molecule and a the area occupied. This has to be modified when the molecule is a member of a dilute gas. Suppose N molecules occupy an area A, then the partititon function for the N molecules is

$$F = \frac{(f_A)^N}{N!} = \left[\frac{2\pi mkT}{h^2}A\right]^N \frac{1}{N!}.$$

Using Stirling's approximation the partition function for a single molecule becomes

$$f = F^{1/N} = \frac{2\pi mkT}{h^2}\frac{Ae}{N} = \frac{2\pi mkT}{h^2}ae.$$

Combination of the partition function for the N molecules and the usual equation for the entropy in terms of partition functions

$$S = Nk\left\{\ln f + T\frac{\partial \ln f}{\partial T}\right\} \tag{2}$$

leads to an expression for the two-dimensional entropy of translation

$$ {}_2S_{\text{trans.}} = R \ln MTa + 65{\cdot}80. \tag{3}$$

If the molecule is free to rotate in n independent ways the rotational partition function given by Halford (1934) is

$$f_{\text{rot.}} = \frac{1}{\pi\sigma}\left\{\frac{8\pi^3(I_A^a I_B^b \dots I_G^g)^{1/n}\, kT}{h^2}\right\}^{\frac{1}{2}n},$$

where $a+b+\dots+g = n$, I_A, I_B, etc., are the moments of inertia and σ is the symmetry factor. Hence

$$S_{\text{rot.}} = R\left[\ln \frac{1}{\pi\sigma}\left\{\frac{8\pi^3(I_A^a I_B^b \dots I_G^g)^{1/n}\, kT}{h^2}\right\}^{\frac{1}{2}n} + \frac{n}{2}\right]. \tag{4}$$

In these expressions the degeneracy of the lowest energy state has been omitted because it will be unaffected by the process of adsorption. The symmetry number has been retained because it depends on the ability of the molecule to rotate and may alter on adsorption.

THE ENTROPY OF ADSORPTION OF BENZENE

Because of the rigid nature of the nucleus, benzene and its simple derivatives have received considerable attention in recent years from the point of view of statistical thermodynamics. Pitzer & Scott (1943) calculated the entropy of benzene to be 64·43 entropy units, compared with an experimental value of 64·39 e.u. at 25·0° C. They used the symmetry number 12, and their moments of inertia were based on the C-C distance 1·39 A and the C-H distance 1·08 A. The vibration entropy was found to be 4·70 e.u. and the entropy due to translation and overall rotation 59·73 e.u. The entropy of translation and rotation for a gram-molecule of benzene confined to a plane at the standard concentration of 1 mol./22·53T A, and under different amounts of rotational freedom, will now be calculated. The translational entropy is given by equation (3). The form of the equation (4) used for the vapour phase rotation is

$$S_{\text{rot.}} = R\left[\ln \frac{1}{\pi\sigma}\left\{\frac{8\pi^3(I_A I_B^2)^{\frac{1}{3}}\, kT}{h^2}\right\}^{\frac{3}{2}} + \frac{3}{2}\right].$$

The vibrational changes are ignored as a first approximation; this is possible because the contributions to the entropy of the vibrational degrees of freedom are small at room temperature. The most natural assumption to make about a benzene molecule, when adsorbed flat on a surface, is that rotation can occur in the plane of the ring but not about either of the other two axes. If this is so the symmetry number will become 6 and the rotational entropy will be

$$S_{\text{rot.}} = R\left[\ln \frac{1}{\pi\sigma}\left\{\frac{8\pi^3 I_A kT}{h^2}\right\}^{\frac{1}{2}} + \frac{1}{2}\right].$$

In table 1 the possible assumptions, the corresponding translational and rotational entropies of the 'adsorbed phase', i.e. the molecules confined to a plane in a force-free field, and the resulting decrease in entropy from the three-dimensional phase are given. The constants used are

$$M = 78, \quad I_A = 2\cdot906 \times 10^{-38}\,\text{g.cm.}^2, \quad I_B = I_C = 1\cdot453 \times 10^{-38}\,\text{g.cm.}^2;$$

these values agree with those used by Pitzer & Scott (1943).

TABLE 1. ENTROPY OF ADSORPTION OF BENZENE. TEMP. = 323·1° K

description of phase	entropy (e.u.) trans.	rot.	total	entropy loss
three-dimensional	39·4	21·0	60·4	0·0
two-dimensional without rotation	30·4	0·0	30·4	30·0
two-dimensional with rotation in plane of ring	30·4	4·0	34·4	26·0
two-dimensional with rotation in plane of ring and about one of the axes in the plane of the ring	30·4	11·8	42·2	18·2
two-dimensional with all three rotations	30·4	21·0	51·4	9·0

experimental loss on adsorption 25·2 ± 0·5 e.u.

The agreement between the experimental entropy of adsorption and that calculated for molecules rotating only in the plane of the ring is striking, but no doubt partly fortuitous because the new vibrations replacing the lost modes of translation and rotation have been neglected. Again, it is possible that the two translational degrees of freedom are only partially free and that the rotations are only partially hindered.

THE ENTROPY OF ADSORPTION OF TOLUENE

Toluene like benzene can be treated accurately by the methods of statistical thermodynamics. The only uncertainty is the hindrance to free rotation of the methyl group. Pitzer & Scott (1943) obtain the theoretical value for the entropy of the vapour at 1 atm. and 25° C of 76·73 e.u., comparing favourably with the experimental 76·44 e.u. The difference, they suggest, is due to an energy barrier of 500 cal. in the rotation of the methyl group. The symmetry number is 6, of which 3 will be associated with the methyl group and 2 with rotation about the axis lying in the plane of the ring and passing through the carbon atom of the methyl group. The axes of the rotations are shown in figure 1, and the entropies calculated for the substance confined in a force-free field under different conditions are given in table 2. The values used for the moments of inertia are those given by Pitzer & Scott (1943): $I_A = 4\cdot94 \times 10^{-38}\,\text{g.cm.}^2$, $I_B = 1\cdot506 \times 10^{-38}\,\text{g.cm.}^2$, $I_C = 3\cdot50 \times 10^{-38}\,\text{g.cm.}^2$ and $I_{\text{rot.}} = 0\cdot0508 \times 10^{-38}\,\text{g.cm.}^2$.

One might have expected that, like benzene, the allowance of two-dimensional translation, and rotation about the axis AA, would represent the state of the toluene molecule when adsorbed 'flat'. However, the adsorption appears to be much stronger (cf. the heats of adsorption, benzene 16·0 and toluene 21·2 kcal./g.mol.) and limits the freedom of the molecules to a greater extent. The denial of all rotation

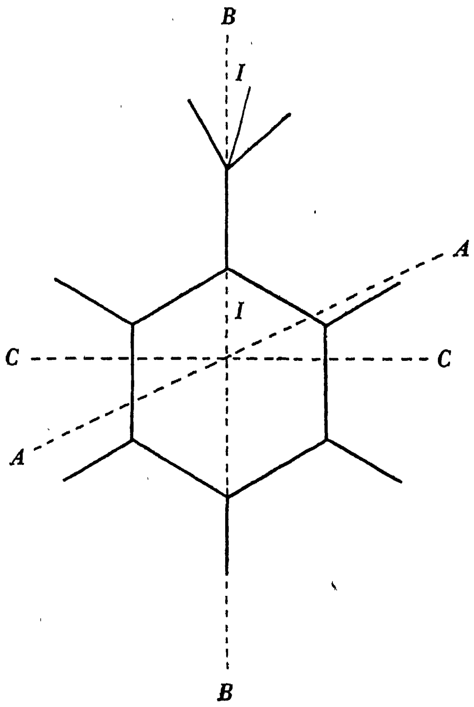


FIGURE 1

TABLE 2. THE ENTROPY OF ADSORPTION OF TOLUENE. TEMP. 310·6° K

description of phase	symmetry number	entropy (e.u.) trans.	external rotation	internal rotation	total	loss
three-dimensional	6	39·7	25·8	3·6	69·1	0·0
two-dimensional:						
no rotation	1	30·6	0·0	0·0	30·6	38·5
rotation about *AA*	1	30·6	8·1	0·0	38·7	30·4
rotation about *BB*	2	30·6	5·5	0·0	36·1	33·0
rotation about *AA* and *II*	3	30·6	8·1	3·6	42·3	26·8
rotation about *BB* and *II*	6	30·6	5·5	3·6	39·7	29·4
rotation about *AA*, *BB* and *II*	6	30·6	15·8	3·6	50·0	19·1
rotation about *AA*, *CC* and *II*	3	30·6	18·1	3·6	52·3	16·8
rotation about *BB*, *CC* and *II*	6	30·6	15·5	3·6	49·7	19·4
two-dimensional, all rotations	6	30·6	25·8	3·6	60·0	9·1

experimental loss on adsorption first phase 39·6 ± 0·9 e.u.
second phase 15·2 ± 1·6 e.u.

would fit the observed entropy of the first phase, but it is unlikely that all rotation will cease before the translation is considerably affected. The calculations indicate that the methyl group of toluene acts as an anchor restricting translations and rotations and giving rise to an additional heat of adsorption of 5 kcal./g.mol. from the benzene value.

When the toluene is adsorbed vertically (e.g. at high surface pressures when the co-area is 23 A^2/mol.) the molecules possess greater freedom. The rotations about BB and II would be expected, but something else is required as well to account for the low entropy of adsorption. The rotation about CC probably remains because this involves the passage of the molecule through the more stable position of 'flat' adsorption. The barrier to this rotation will be caused only by the pressure of the other molecules, which will be small under the conditions of the standard state. The rotation about AA for a 'vertically' adsorbed molecule could only take place if the molecule passed through a position of 'edge' adsorption, which would be of higher energy. In choosing which of the conditions outlined in table 2 appear to fit the experimental values, the neglected vibrations must be remembered. It is better to choose conditions involving a greater decrease of entropy on adsorption than the experimental, as the difference will be partly compensated by the additional vibrational entropy.

THE ENTROPY OF ADSORPTION OF *n*-HEPTANE

The entropy of *n*-heptane has been calculated in two ways, both giving agreement with the experimental value of 101·3 e.u. at 298·1° K. Pitzer (1940) assumed that the molecules are randomly linked under the restraints imposed by the potential barriers to the internal rotations. Using standard frequencies for the bending and stretching of the carbon-carbon link, and for the motions of the hydrogen atoms he calculated the entropy to be 101·54 e.u. at 298·1° K. The other method of calculating the entropy of *n*-heptane was due to Aston, Isserow, Szasz & Kennedy (1944). They deduced from the calculation of the potential energy of the ethyl groups in *n*-butane that the straight chain position was the most stable form, and concluded that most of the smaller normal paraffins would be in the straight chain position in the gas phase. By choosing the carbon-carbon bending frequencies to give the observed entropy at the boiling-point under atmospheric pressure, and choosing suitable rocking frequencies for the methyl group to fit the experimental heat-capacity curves, they fitted the straight chain model to the experimental data.

It is clearly impractical to calculate the effect of all the possible restrictions in the surface phase on the entropy of *n*-heptane. However, it is possible, using the knowledge outlined above about the entropy, to confirm the picture of the adsorbed heptane molecules presented in the previous paper. It was suggested from considerations of co-area that the heptane molecules were partially curled up on the surface and that four or five of the carbon atoms were in contact with the mercury at any one time. The experimental entropy of adsorption was 15·3 ± 0·7 e.u. at 310·6° K. The calculations will, however, be made for heptane at 298·1° K. Equation (1) gives 39·7 e.u. for translation in the gas phase and equation (3) gives 30·6 e.u. for two-dimensional translation of the adsorbed molecules. This difference accounts for 9·1 e.u. of the experimental loss of 15·3 e.u. The remaining 6·2 e.u. must be connected with the change in the rotational freedom of the molecules on adsorption. Taking the values suggested by Aston *et al.* (1944) for the vibrational frequencies of *n*-heptane, namely, carbon-carbon bending 165 cm.$^{-1}$ (5), carbon-carbon stretching 1000 cm.$^{-1}$ (6), hydrogen rocking 1070 cm.$^{-1}$ (19), hydrogen bending 1400 cm.$^{-1}$ (11), and hydrogen stretching 3000 cm.$^{-1}$ (16), then the vibrational entropy is 14·6 e.u. at 298·1° K. Subtraction of this and the translation entropy from the experimental value of 101·3 e.u. gives the entropy due to overall and internal rotation of *n*-heptane as 47·0 e.u. A similar calculation for *n*-heptane gives 41·9 e.u. as the rotational entropy. Both these calculations are based on a straight-chain form for the hydrocarbon because the vibrational frequencies were assigned assuming that model. The increase of 5·1 e.u. for rotation between hexane and heptane corresponds to two factors: first, an increase in the moments of inertia for overall rotation; and second, the new restricted rotation or torsional oscillation about the extra bond. The second of these factors is the more important, as the entropy depends only on the logarithm of the half-power of the moment of inertia. The similarity between 5·1 and 6·2 means that one possible picture of adsorbed heptane is that the molecule is doubled up, rotation about one of the central bonds being denied, but the rest of the molecule is as free to oscillate as in the gas phase. Although this is only one of many possible pictures of the adsorbed molecules, it shows that the entropy of adsorption is consistent with partial curling of the heptane on adsorption. The overall rotations appear to be almost unhindered, as most of the 6·2 e.u. must be connected with restrictions on internal rotation, which means that each molecule must be able to roll freely on the surface like a ball, or in other words must possess something approaching spherical symmetry, or the adsorption for one orientation would be stronger than for the others as in the case of benzene.

Another method of obtaining the change in entropy when the gaseous heptane molecule is converted to a spherical molecule on the surface is afforded by Huggins (1940). He gives general equations for the entropy of hydrocarbon molecules both in the randomly linked state (5) and when rolled into a spherical mass (6):

$$S = 54 - 3k_{ir} + 9\ln n + (1{\cdot}8 + k_{ir})\,n, \tag{5}$$

$$S = 52 - 3k_{ir} + 8\ln n + (1{\cdot}8 + k_{ir})\,n, \tag{6}$$

where k_{ir} is a constant giving the entropy of internal rotation per carbon and n is the number of carbon atoms in the chain. Although his results only apply strictly to large molecules, he shows that the relations are obeyed approximately for the smaller molecules. If $n = 7$ is used, the difference between expressions (5) and (6) is 3·9 e.u. Thus once again the experimental decrease of 6·2 e.u. could be attributed to the rolling of the molecules into balls.

THE HEAT ADSORPTION OF BENZENE

The experimental heat of adsorption of benzene was found to be 16·0 ± 0·2 kcal./g.mol. at 50° C. This must be corrected to the heat of adsorption at the absolute zero. From the entropy considerations it was suggested that the molecules have lost one degree of translational, two of rotational, and gained three of vibrational freedom on adsorption. The specific heat for the translations and rotations will be 1 cal./degree/g.mol. down to low temperatures, whereas, although the vibrational contributions may be as much as 1 cal./degree/g.mol. at room temperature, the value will fall sharply with temperature. The second term on the right-hand side of Kirchhoff's equation (7) will be about 600 cal. taking 2/3 cal./degree/g.mol. for the average alteration for each of the three degrees of freedom:

$$\Delta H_T = \Delta H_0 + \int_0^T (\Delta Cp)\, dT \tag{7}$$

The heat of adsorption at the absolute zero is therefore 15·4 kcal./g.mol.

ATTRACTIVE POTENTIAL

The attractive energy liberated by the approach of a molecule to a metal surface may be calculated from a number of equations, not always in good agreement. If the metal is treated as a covalently linked structure, the attractive potential may be obtained from either the London (1930; Born & Mayer 1932) equation (8) or the Kirkwood (1932) equation (9):

$$U = -\tfrac{9}{4}\alpha_1\alpha_2 \frac{h\nu_1}{h\nu_1 + h\nu_2}\frac{1}{r^6}, \tag{8}$$

$$U = -6mc^2 \frac{\alpha_1\alpha_2}{\alpha_1/\chi_1 + \alpha_2/\chi_2}\frac{1}{r^6}, \tag{9}$$

where α_1, α_2 are polarizabilities, χ_1 and χ_2 are diamagnetic susceptibilities and $h\nu_1$ and $h\nu_2$ ionization potentials of the molecule and the metal. If this method is used, the interaction between the adsorbed molecule and each metal atom is considered and summation has to be made over all the metal atoms. However, there are now a number of equations for the attractive potential near a metal surface considering the metal as a uniform sea of electrons. All apply only to spherically symmetrical molecules, but benzene may be assumed to contain six spherically symmetrical units.

The simplest of these equations, due to Lennard-Jones (1932), gives the energy resulting from the image force experienced by a molecule near a perfectly conducting surface, i.e. the metal is assumed to be completely polarizable and to reproduce instantaneously an image of any charge distribution. The equation (10) thus gives the maximum interaction that would be expected:

$$U = \frac{-mc^2\chi}{2Nr^3}, \tag{10}$$

where m = mass of electron, c = velocity of light, r = distance from surface and χ is the magnetic susceptibility of the benzene. Using $m = 9{\cdot}11 \times 10^{-28}$ g., $c = 3 \times 10^{10}$ cm./sec. and $\frac{\chi}{N} = 93{\cdot}3 \times 10^{-30}$ c.g.s. units, equation (10) gave the '$L-J$' curve shown in figure 2.

Bardeen (1940) treated the metal as well as the molecule quantum mechanically, giving the expression (11):

$$U = -\frac{mc^2\chi}{2Nr^3}\frac{\dfrac{C_Be^2}{2r_sh\nu_0}}{1+\dfrac{C_Be^2}{2r_sh\nu_0}}, \tag{11}$$

where C_B is a constant $\doteqdot 2{\cdot}6$, r_s is the radius of a sphere containing one metal electron, and $h\nu_0$ is the characteristic energy of the benzene, which is approximately equal to the ionization potential. Assuming two electrons per metal atom, i.e. the value of r_s is 1·432 A and the value of 9·19 electron volts (Price & Wood 1935) for $h\nu_0$, the calculated '*B*' curve in figure 2 was obtained.

Prosen & Sachs (1943) gave an entirely different type of expression, depending not on r^{-3} but on $r^{-2}\log r$. Their equation was based on perturbation theory which applies only if the molecules are sufficiently far away from the surface for there to be little overlapping of the wave functions of the molecule and the mercury. The Fermi degeneracy of the electron gas was taken into account, but the Coulombic interaction between the electrons in the metal was neglected. They suggest that the expression (12)

$$U = -\frac{\alpha e^2k_m^2}{8\pi^2}\frac{\ln 2k_mr}{r^2}, \tag{12}$$

where α is the polarizability of the molecule, and k_m is given by

$$k_m^3 = 3\pi\rho,$$

where ρ is the density of electrons in the metal, will break down as soon as the calculated interaction is as large as the Lennard-Jones interaction. However, they point out that the treatments used by Bardeen and by Margenau & Pollard (1941) (see below) only apply to large values of r. They conclude that the Lennard-Jones expression probably sets a maximum value to the attractive potential over a large range of r. Taking $\rho = 8{\cdot}136\times10^{22}$ electrons/c.c. and the mean polarizability of benzene $= 10{\cdot}3\times10^{-24}$ c.c. the '*P.S.*' curve of figure 2 was obtained.

The last method of obtaining the attractive potential was due to Margenau & Pollard (1941), who considered the metal surface as small elements, evaluating the interaction between the dipole field of each element and the molecule. They gave

$$U = -\left[\frac{e^2h^2}{m}A(\nu_0)\frac{f}{\nu_0}+\frac{C_Be^2\alpha}{2r_s}\right]\frac{1}{8r^3}, \tag{13}$$

where the terms have the same meanings as above and $A(\nu_0)$ is the polarizability for the frequency ν_0, and is given by

$$A(\nu_0) = \frac{-\rho e^2}{4\pi^2m\nu_0^2},$$

and f is given by

$$f = \frac{m\alpha\nu_0^2}{e^2}.$$

As can be seen from figure 2, equation (13) (the '*M.P.*' curve) agrees well with the Bardeen equation (11). This does not mean that they are more reliable than the others, but confirms the suggestion of Prosen & Sachs (1943) that they both hold and break down over the same range of values of r. In all cases for the attractive potential it is possible to use the polarizability or diamagnetic susceptibility of the complete benzene molecule, because these quantities have appeared linearly in the equations.

REPULSIVE POTENTIAL

One method of obtaining the net energy of interaction is to assume an equilibrium distance and that repulsive potential is a definite fraction of the attractive potential at that distance. A second method assumes only the equilibrium distance and the form of the repulsive potential, and using the condition that the two forces balance at that distance the total interaction may be evaluated. Both methods involve a number of arbitrary assumptions. However, Pollard (1941) has given an equation for the repulsive energy experienced by molecules possessing 1s electrons near a metal surface. This equation may be used to give an approximate repulsive potential for benzene, by considering the electrons in benzene as 1s electrons of the same polarizability. The expression given by Pollard (1941) was

$$U = \frac{e^2[p(\rho_s)\,G(D) + q(\rho_s)\,(p(\rho_s)-1)\,F(D)]}{[1-p(\rho_s)\,F(D)]}, \tag{14}$$

where

$$F(D) = \frac{1}{2}\left(1+\frac{D}{a}\right)e^{-2D/a},$$

$$G(D) = \left(\frac{1}{2a}\right)e^{-2D/a},$$

$$p(\rho_s) = 3\left[\frac{\rho_s}{a}\cosh\left(\frac{\rho_s}{a}\right) - \sinh\left(\frac{\rho_s}{a}\right)\right]\Big/\left(\frac{\rho_s}{a}\right)^3,$$

$$q(\rho_s) = \frac{3}{a}\left[\cosh\left(\frac{\rho_s}{a}\right) - 1\right]\Big/\left(\frac{\rho_s}{a}\right)^3,$$

and

$$\rho_s = 2^{\frac{1}{3}} r_s,$$

where r_s is the radius of a sphere containing one metal electron of a given spin (assumed to be one electron per metal atom), D is the distance from the surface and a is given by

$$a = \left[\frac{h^2\alpha}{8me^2}\right]^{\frac{1}{4}},$$

where the symbols have the usual meanings.

The thirty valency electrons in benzene may be considered in three groups, the twelve electrons forming C-H bonds, the twelve forming the C-C bonds and the six π electrons. The description of the benzene molecule given by Coulson (1941) indicates that most of the repulsive potential will be due to the interaction between the surface and the π electrons, as they project considerably farther from the molecule than the others. The repulsive potential for each of these groups of electrons will be calculated for a series of values of D. The necessary information about the longitudinal and transverse polarizability of the different bonds was given by Denbigh (1940). Assuming one electron per metal atom the value of r_s is $2{\cdot}274 \times 10^{-8}$ cm.

The twelve electrons forming the C-H bonds will now be considered. The transverse polarizability of the C-H bond is $5{\cdot}8 \times 10^{-25}$ c.c., i.e. $2{\cdot}9 \times 10^{-25}$ c.c./electron, giving a value of 'a' of $0{\cdot}492 \times 10^{-8}$ cm. from equation (15). Now the electronic radii (i.e. the distance from the nucleus at which the radial charge density is a maximum) given by Slater (1939) are for the hydrogen 1s electron 0·53 A, for the carbon 2s 0·67 A, for the carbon 2p 0·66 A and for the carbon 1s (which do not have to be considered) 0·090 A. Thus the two electrons which form the C-H bond have radii 0·53 and 0·66 A, and the bond ellipsoid is being considered as two spheres with maximum charge density at 0·492 A, a reasonable value in view of the concentration of the electrons along the bond. The repulsive potentials obtained by using $a = 0{\cdot}492$ A and a series of values of D in expression (14) are given in table 3.

TABLE 3. THE REPULSIVE POTENTIAL FOR BENZENE-MERCURY

distance $\times 10^8$ cm.	12 C-H electrons kcal./g.mol.	12 C-C electrons kcal./g.mol.	6π electrons kcal./g.mol.	Σ repulsive kcal./g.mol.
2·0	48·96	22·32	71·6	142·9
2·5	6·685	0·516	13·14	20·34
3·0	1·123	0·011	2·392	3·526
3·5	0·166	0·000	0·427	0·593
4·0	0·024	—	0·060	0·084
4·5	0·003	—	0·013	0·016

Denbigh (1940) gave the transverse polarizability of each bond in the benzene ring as $4{\cdot}8 \times 10^{-25}$ c.c. This corresponds to three electrons, the contribution of one being much greater than the other two. The transverse polarizability of a single C-C is $0{\cdot}2 \times 10^{-25}$ c.c., i.e. $0{\cdot}1 \times 10^{-25}$ c.c./electron. Taking double this value for the two electrons in the benzene bond 'a' is found to be 0·252 A, a value in keeping with a very flattened ellipsoid from two electrons originally of size 0·66 A. Despite the fact that the transverse polarizability has been taken as double that of the ordinary C-C bond, table 3 shows that the repulsive potential due to these twelve electrons is small and the approximation will not affect the total potential greatly.

Having assigned $0{\cdot}2 \times 10^{-25}$ c.c. of the $4{\cdot}8 \times 10^{-25}$ c.c. to each of the two bond forming electrons, $4{\cdot}4 \times 10^{-25}$ c.c. remains for each of the π electrons. The calculated value of 'a' is 0·546 A, again in agreement with the Slater value for the unbonded 2p electron of carbon with the Coulson picture of slight interaction between these electrons in the benzene ring. In the calculations described above the electrons have been assumed to be spherically symmetrical, i.e. they have been replaced by 1s electrons of equivalent size. This has been permissible because ellipsoids have been divided into two spheres, but in the present case the π electrons are concentrated along one axis which projects out from the plane of the ring. The repulsion will be greater, therefore, than the value calculated for a spherical electron, and it is proposed to consider each of the π electrons as equivalent to three spherically symmetrical electrons of radius 0·546 A to allow for the fact that the electrons are concentrated in one direction. From table 3 it can be seen that the π electrons give rise to the majority of the repulsive potential.

The logarithm of the total repulsive potential is nearly linear with distance (figure 3). Potential energy curves (figure 4) are constructed from figure 3 and from two of the curves for the attractive energy given in figure 2. The Lennard-Jones curve is used to give a maximum limit to the energy of adsorption, 18·4 kcal./g.mol. at 2·7 A, and the Margenau & Pollard, a lower limit, 10·2 kcal./g.mol. at 2·95 A. These results bracket the experimental value of 15·4 kcal./g.mol., indicating that the various relations are approximately correct. It would be possible to draw a smoothed curve following the Prosen & Sachs relation at small distances and gradually changing over to the Bardeen or the Margenau & Pollard at greater distances, which would give a value near the experimental value. This confirms the suggestions of Prosen & Sachs (1942) about the ranges in which the various equations hold, namely, the Prosen & Sachs relation at small distances from the surface, and

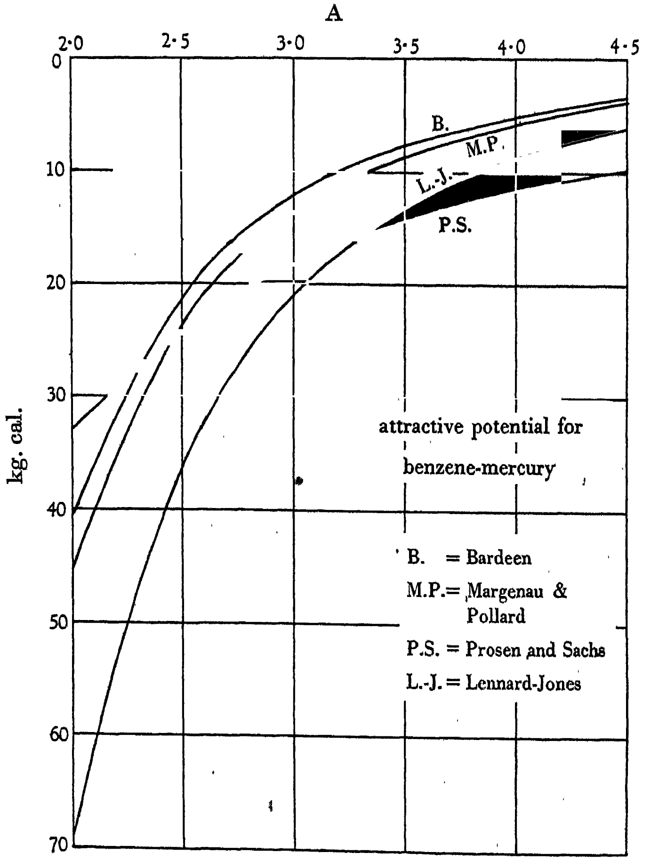


FIGURE 2

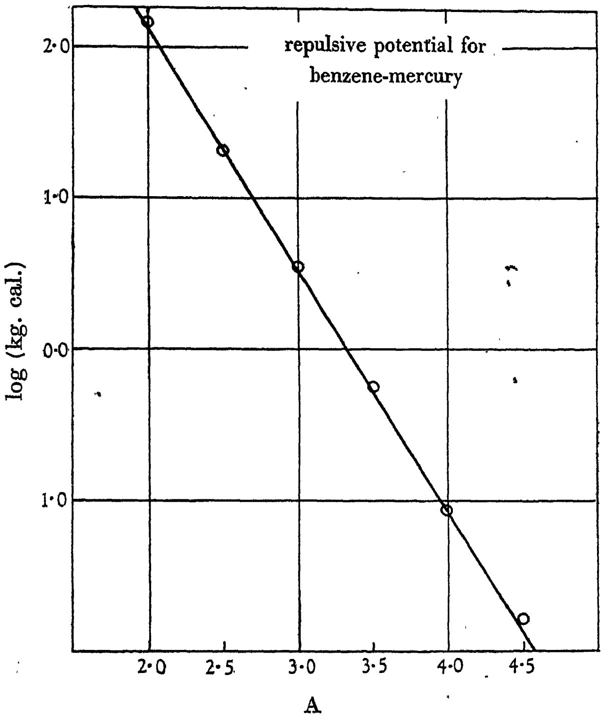


FIGURE 3

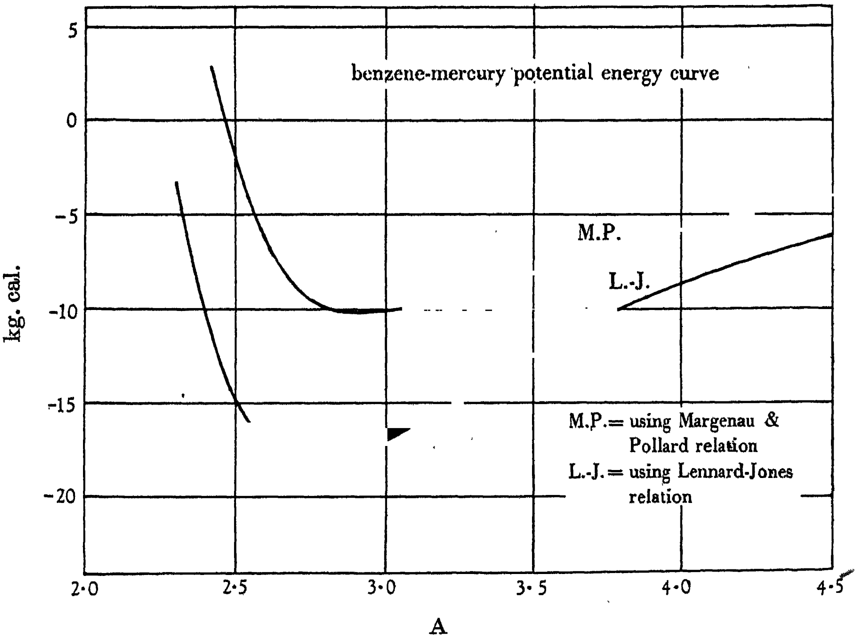


FIGURE 4

the Bardeen and the Margenau & Pollard at greater distances, with the Lennard-Jones as an upper limit at all distances. The values for the equilibrium distances appear to be small, as Pauling (1942) gives the van der Waals half-thickness of an aromatic molecule as 1·85 A and the radius of a mercury atom is about 1·5 A. Most of the relations treat the mercury as a uniform structure with electrons in cubical elements and presumably count the surface as starting from the centre of the first plane cubes which may be nearer the adsorbed molecule than the first plane of mercury atoms.

The greatest objection to the type of relation which has been used above is that no indication is obtained of the periodic nature of the energy at the mercury surface which will govern the whole question of surface mobility. For this reason it would have been more satisfactory to use the equations of London and Slater & Kirkwood for the interaction of molecules and covalently linked structures. The interaction of each spherically symmetrical part of the molecule and each of the mercury atoms within a certain distance would require summation. It is doubtful if the labour of such a calculation would be justified in view of the assumptions which would have to be made about the repulsive forces. It is felt that the calculations which have been carried out for the adsorption of benzene are sufficient to indicate that the values for the heats of adsorption of the non-polar substances adsorbed on mercury are due to the van der Waals dispersive forces. The calculations afford some evidence as to the regions in which the various relationships for the van der Waals dispersive energy can be applied accurately. There appears to be no equation suitable for the required distance of some 3 A from the surface, the Prosen & Sachs giving correct results slightly nearer the surface, and the Bardeen and the Margenau & Pollard equations holding farther from the surface. A number of approximations have had to be made, especially in the calculation of the repulsive potential. However, there is still a large gap to be bridged before accurate equations, as yet only applicable to small spherically symmetrical molecules under well-defined conditions, can be applied to such a simple chemical system as the adsorption of benzene on mercury.

I have pleasure in thanking Professor Eric K. Rideal, F.R.S., for his interest and encouragement in this work which was financed by the Ministry of Aircraft Production. I am also grateful to Mr A. H. Wilson, F.R.S., for some advice on the theoretical calculation of energies of adsorption, and to Dr A. D. Walsh for helpful discussion.

REFERENCES

Aston, Isserow, Szasz & Kennedy 1944 *J. Chem. Phys.* **12**, 336.
Bardeen 1940 *Phys. Rev.* **58**, 727.
Born & Mayer 1932 *Z. Phys.* **75**, 1.
Coulson 1941 *Proc. Roy. Soc. Edinb.* **61**, 115.
Denbigh 1940 *Trans. Faraday Soc.* **36**, 936.
Halford 1934 *J. Chem. Phys.* **2**, 694.
Huggins 1940 *J. Chem. Phys.* **8**, 181.

Kirkwood 1932 *Phys. Z.* **33**, 57.
Lennard-Jones 1932 *Trans. Faraday Soc.* **28**, 333.
London 1930 *Z. phys. Chem.* B, **11**, 222.
Margenau & Pollard 1941 *Phys. Rev.* **60**, 128.
Pauling 1942 *The nature of the chemical bond.* Cornell University Press.
Pitzer 1940 *J. Chem. Phys.* **8**, 711.
Pitzer & Scott 1943 *J. Amer. Chem. Soc.* **65**, 803.
Pollard 1941 *Phys. Rev.* **60**, 578.
Price & Wood 1935 *J. Chem. Phys.* **3**, 439.
Prosen & Sachs 1942 *Phys. Rev.* **61**, 65.
Slater 1939 *Introduction to chemical physics.* McGraw-Hill.

Two-dimensional boundary-value problems of elasticity

By Bibhutibhusan Sen

(*Communicated by R. V. Southwell, F.R.S.—Received* 21 *November* 1941)

A new method of solving two-dimensional problems of elasticity is developed in this paper. Airy's stress function is not used; but the stress equations of equilibrium are so transformed with the help of consistency equations as to yield simple solutions when the values of the stresses on the boundary are given. The boundaries discussed in this paper are the straight edges of a semi-infinite plate and the circular boundaries of a thin disk. Action of forces and couples in the interior of the plates is chiefly considered.

INTRODUCTION

Plane problems of elasticity are usually solved with the help of Airy's stress function which satisfies a differential equation of the fourth order. In two papers published by the author* (1938, 1939), it has been shown that the stresses in certain two-dimensional problems of elasticity can be directly obtained with much advantage from the stress equations and consistency relations. Considering each problem to be one of generalized plane stress, it has been shown that the stress distribution in a circular plate or a semi-infinite plate with a straight edge can be easily found for prescribed values of stresses on the boundary. By simple and suitable modifications, the different cases are reduced to boundary-value problems of the theory of potential. It is a known fact that the two-dimensional potential problems are much simplified if the functions of complex variable are used. It has been found that the use of these functions also facilitates the solution of two-dimensional boundary-value problems of elasticity. The object of this paper is to solve, with the help of complex functions, the problems in which the boundary is free while a single force or a couple (with its axis normal to the plate) operates at any point inside a thin plate.

---

* These two papers will be referred to as Paper I and Paper II respectively.

The general outline of the method is as follows:

At first consider the thin plate to be of indefinite length in all directions with faces parallel to the x-y plane. A force or a couple acting at a point in such a plate will produce stress components $\widehat{xx}_1$, $\widehat{xy}_1$ and $\widehat{yy}_1$ with singularities at this point. The values of these components are already known and are given in treatises on the theory of elasticity. If these values for a plate having a circular boundary or a semi-infinite plate having a straight edge now be assumed (with the point of application of the force or the couple in the interior of the plate), the boundary will not necessarily be free from stress. However, the values of the *surface* tractions $\widehat{x\nu}$, $\widehat{y\nu}$, can be calculated from the stress components $\widehat{xx}_1$, $\widehat{xy}_1$ and $\widehat{yy}_1$. To nullify these surface tractions on the boundary it will be necessary to superpose another system of stress components $\widehat{xx}_2$, $\widehat{xy}_2$, $\widehat{yy}_2$ [with singularities at an outside point] such that the surface tractions $\widehat{x\nu}_2$, $\widehat{y\nu}_2$ calculated from them give us $\widehat{x\nu}_2+\widehat{x\nu}_1 = 0$ and $\widehat{y\nu}_2+\widehat{y\nu}_1 = 0$ on the boundary. In order that the added system may be physically possible, the stress components $\widehat{xx}_2$, $\widehat{xy}_2$, $\widehat{yy}_2$ must satisfy the equations of equilibrium and the consistency relations. These stresses being found, the resultant stress components will be given by

$$\widehat{xx} = \widehat{xx}_1+\widehat{xx}_2, \quad \widehat{xy} = \widehat{xy}_1+\widehat{xy}_2 \quad \text{and} \quad \widehat{yy} = \widehat{yy}_1+\widehat{yy}_2.$$

It can be easily seen that this method is comparable to that of images which is widely used in various branches of mathematical physics. To illustrate the application of the method, the following problems have been selected for solution.

Section I

(*a*) The problem of the stress distribution due to a single force acting in the interior of an isotropic plate with a straight edge.

(*b*) The problem of the stress distribution in an isotropic semi-infinite plate with a straight edge due to a single couple operating at a point inside it.

(*c*) The problem of the stress distribution due to a point of pressure situated inside a plate as described above.

Section II

(*a*) The problem of the stress distribution due to an isolated couple operating at a point inside a semi-infinite plate of certain non-isotropic material with a straight edge.

(*b*) The problem of the stress distribution due to a point of pressure situated inside a plate of the above type.

Section III

(*a*) The problem of the stress distribution due to couples operating in the interior of an isotropic circular plate.

(*b*) The problem of the stress distribution due to forces acting in the interior of an isotropic circular plate.

Problems I (*a*) and I (*b*) were solved in a different manner by Melan (1932) and Ghosh (1937) respectively. But the remaining results, and especially the method used in this paper, are believed to be new.

SECTION I

(*a*) *Stresses due to a single force acting in the interior of an isotropic semi-infinite plate in the direction normal to the straight boundary*

Let the axis of y coincide with the straight boundary of the semi-infinite plate of unit thickness, the axis of x being drawn into the plate perpendicular to the edge. Suppose a force of magnitude T acts at the point $(c, 0)$ along the axis of x. If the plate had extended to an indefinite length in all directions, the stress components due to the force would have been (Timoshenko 1932)

$$\left.\begin{aligned} \widehat{xx}_1 &= \frac{T}{4\pi}\left[-\frac{(1-\sigma)(x-c)}{(x-c)^2+y^2}-\frac{2(1+\sigma)(x-c)^3}{\{(x-c)^2+y^2\}^2}\right], \\ \widehat{yy}_1 &= \frac{T}{4\pi}\left[\frac{(1-\sigma)(x-c)}{(x-c)^2+y^2}-\frac{2(1+\sigma)y^2(x-c)}{\{(x-c)^2+y^2\}^2}\right], \\ \widehat{xy}_1 &= -\frac{T}{4\pi}\left[\frac{(1-\sigma)y}{(x-c)^2+y^2}+\frac{2(1+\sigma)(x-c)^2y}{\{(x-c)^2+y^2\}^2}\right], \end{aligned}\right\} \tag{1·1}$$

where σ stands for Poisson's ratio. On the line $x = 0$, then

$$\left.\begin{aligned} [\widehat{xx}_1]_{x=0} &= \frac{T}{4\pi}\left[\frac{2c}{c^2+y^2}+\frac{(1+\sigma)c(c^2-y^2)}{(c^2+y^2)^2}\right], \\ [\widehat{xy}_1]_{x=0} &= \frac{T}{4\pi}\left[-\frac{(1-\sigma)y}{c^2+y^2}-\frac{2(1+\sigma)c^2y}{(c^2+y^2)^2}\right]. \end{aligned}\right\} \tag{1·2}$$

In order that the bounding edge may be free from stress, it will be necessary to superpose on the stresses given in (1·1) stresses $\widehat{xx}_2$, $\widehat{xy}_2$, $\widehat{yy}_2$ which have no singularities inside the plate and which are such that

$$\left.\begin{aligned} [\widehat{xx}_2]_{x=0} &= -[\widehat{xx}_1]_{x=0} = -\frac{T}{4\pi}\left[\frac{2c}{c^2+y^2}+\frac{(1+\sigma)c(c^2+y^2)}{(c^2+y^2)^2}\right], \\ [\widehat{xy}_2]_{x=0} &= -[\widehat{xy}_1]_{x=0} = \frac{T}{4\pi}\left[\frac{(1-\sigma)y}{c^2+y^2}+\frac{2(1+\sigma)c^2y}{(c^2+y^2)^2}\right]. \end{aligned}\right\} \tag{1·3}$$

If z be written for $x+iy$ and R_e for the real part of a complex function, (1·3) can be put in the form

$$\left.\begin{aligned} [\widehat{xx}_2]_{x=0} &= -\frac{T}{4\pi}\left[R_e\left\{\frac{2}{z+c}+\frac{(1+\sigma)c}{(z+c)^2}\right\}\right]_{x=0} = [R_e f_1(z)]_{x=0}, \\ [\widehat{xy}_2]_{x=0} &= \frac{T}{4\pi}\left[R_e\left\{\frac{(1-\sigma)i}{z+c}+\frac{(1+\sigma)ci}{(z+c)^2}\right\}\right]_{x=0} = [R_e f_2(z)]_{x=0}. \end{aligned}\right\} \tag{1·4}$$

Functions $f_1(z)$ and $f_2(z)$ in the above expressions have singularities at the point $(-c, 0)$ which is outside the plate. It has been shown in an earlier paper [Paper I art. I (*b*)] that if the boundary values of the stresses $\widehat{xx}_2$, $\widehat{xy}_2$ on the line $x = 0$ be given as $[R_e f_1(z)]_{x=0}$ and $[R_e f_2(z)]_{x=0}$ respectively, the stress system is given by

$$\left.\begin{aligned}\widehat{xx}_2 &= R_e[-xf_1'(z)+f_1(z)-ixf_2'(z)],\\ \widehat{xy}_2 &= R_e[-ixf_1'(z)+xf_2'(z)+f_2(z)],\\ \widehat{yy}_2 &= R_e[f_1(z)+xf_1'(z)+2if_2(z)+ixf_2'(z)],\end{aligned}\right\} \tag{1·5}$$

where $f_1'(z)$ and $f_2'(z)$ denote the derivatives of the functions $f_1(z)$ and $f_2(z)$ with respect to z. Hence, from (1·4),

$$f_1(z) = -\frac{T}{4\pi}\left[\frac{2}{z+c}+\frac{(1+\sigma)c}{(z+c)^2}\right], \quad f_2(z) = \frac{T}{4\pi}\left[\frac{(1-\sigma)i}{z+c}+\frac{(1+\sigma)ci}{(z+c)^2}\right]. \tag{1·6}$$

On substituting these values in (1·5), the stress components are obtained as

$$\left.\begin{aligned}\widehat{xx}_2 &= \frac{T}{4\pi}R_e\left[-\frac{2x}{(z+c)^2}-\frac{2(1+\sigma)xc}{(z+c)^3}-\frac{2}{z+c}-\frac{(1+\sigma)c}{(z+c)^2}-\frac{(1-\sigma)x}{(z+c)^2}-\frac{2(1+\sigma)cx}{(z+c)^3}\right],\\ \widehat{xy}_2 &= \frac{T}{4\pi}R_e\left[-\frac{2ix}{(z+c)^2}-\frac{2(1+\sigma)cix}{(z+c)^3}-\frac{(1-\sigma)ix}{(z+c)^2}\right.\\ &\qquad\left.-\frac{2(1+\sigma)cix}{(z+c)^3}+\frac{(1-\sigma)i}{z+c}+\frac{(1+\sigma)ci}{(z+c)^2}\right],\\ \widehat{yy}_2 &= \frac{T}{4\pi}R_e\left[-\frac{2}{z+c}-\frac{(1+\sigma)c}{(z+c)^2}+\frac{2x}{(z+c)^2}+\frac{2(1+\sigma)cx}{(z+c)^3}\right.\\ &\qquad\left.-\frac{2(1-\sigma)}{z+c}-\frac{2(1+\sigma)c}{(z+c)^2}+\frac{(1-\sigma)x}{(z+c)^2}+\frac{2(1+\sigma)cx}{(z+c)^3}\right].\end{aligned}\right\} \tag{1·7}$$

Combining (1·1) and (1·7) and writing r_1^2 for $(x-c)^2+y^2$, and r_2^2 for $(x+c)^2+y^2$, then

$$\begin{aligned}\widehat{xx} &= \widehat{xx}_1+\widehat{xx}_2\\ &= -\frac{T}{4\pi}\left[(1-\sigma)\left\{\frac{x-c}{r_1^2}+\frac{x+c}{r_2^2}+\frac{2x}{r_2^4}[(x+c)^2-y^2]\right\}\right.\\ &\qquad\left.+(1+\sigma)\left\{\frac{2(x-c)^3}{r_1^4}+\frac{2(x+c)^3}{r_2^4}+\frac{4cx(x+c)}{r_2^6}[(x+c)^2-3y^2]\right\}\right],\\ \widehat{xy} &= \widehat{xy}_1+\widehat{xy}_2\\ &= -\frac{Ty}{4\pi}\left[(1-\sigma)\left\{\frac{1}{r_1^2}-\frac{1}{r_2^2}+\frac{4x(x+c)}{r_2^4}\right\}+(1+\sigma)\left\{\frac{2(x-c)^2}{r_1^4}+\frac{2(x^2-c^2)}{r_2^4}+\frac{4cx}{r_2^6}[3(x+c)^2-y^2]\right\}\right],\\ \widehat{yy} &= \widehat{yy}_1+\widehat{yy}_2\\ &= -\frac{T}{4\pi}\left[(1-\sigma)\left\{-\frac{(x-c)}{r_1^2}+\frac{3(x+c)}{r_2^2}-\frac{2x}{r_2^4}[(x+c)^2-y^2]\right\}\right.\\ &\qquad\left.+(1+\sigma)\left\{\frac{2y^2(x-c)}{r_1^4}+\frac{x+c}{r_2^2}-\frac{(x-3c)[(x+c)^2-y^2]}{r_2^4}-\frac{4cx(x+c)[(x+c)^2-3y^2]}{r_2^6}\right\}\right].\end{aligned} \tag{1·8}$$

Proceeding in the same way the stress distribution due to an isolated force acting at an internal point parallel to the straight edge may be found.

(*b*) *Stresses due to a couple acting at a point inside a thin semi-infinite plate with a straight edge*

Let a couple of moment Q with its axis normal to the plate act at the point $(c, 0)$ inside the plate, the axes being taken as shown in figure 1. If such a couple operates at a point $(c, 0)$ inside a plate unlimited in all directions, the stresses produced by it will be given by (Love 1927)

$$\widehat{rr} = \widehat{\theta\theta} = 0 \quad \text{and} \quad \widehat{r\theta} = -\frac{Q}{2\pi r_1^2}, \quad \text{where} \quad r_1^2 = (x-c)^2+y^2.$$

This is equivalent to the system

$$\widehat{xx}_1 = \frac{Q}{\pi}\frac{(x-c)\,y}{[(x-c)^2+y^2]^2}, \quad \widehat{xy}_1 = \frac{Q}{2\pi}\frac{y^2-(x-c)^2}{[(x-c)^2+y^2]^2}, \quad \widehat{yy}_1 = -\frac{Q}{\pi}\frac{(x-c)\,y}{[(x-c)^2+y^2]^2}. \tag{1·9}$$

The boundary edge $x = 0$ will not be free from stress. On this boundary the stresses $\widehat{xx}_1$ and $\widehat{xy}_1$ will be obtained as

$$[\widehat{xx}_1]_{x=0} = -\frac{Q}{\pi}\frac{cy}{(c^2+y^2)^2}, \quad [\widehat{xy}_1]_{x=0} = \frac{Q}{2\pi}\frac{y^2-c^2}{(c^2+y^2)^2}. \tag{1·10}$$

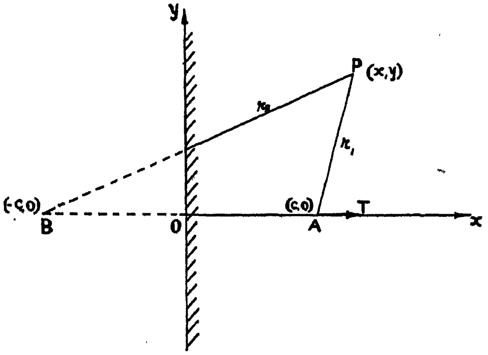


FIGURE 1

To cancel these tractions on the boundary the stress system $\widehat{xx}_2$, $\widehat{xy}_2$, $\widehat{yy}_2$ may be added with singularities at the image point $(-c, 0)$ such that

$$[\widehat{xx}_2]_{x=0} = -[\widehat{xx}_1]_{x=0} = \frac{Q}{\pi}\frac{cy}{(c^2+y^2)^2}, \quad [\widehat{xy}_2]_{x=0} = -[\widehat{xy}_1]_{x=0} = -\frac{Q}{2\pi}\frac{y^2-c^2}{(c^2+y^2)^2}. \tag{1·11}$$

These expressions can be put as

$$[\widehat{xx}_2]_{x=0} = \frac{Q}{2\pi}\left[R_e\frac{i}{(z+c)^2}\right]_{x=0}, \quad [\widehat{xy}_2]_{x=0} = \frac{Q}{2\pi}\left[R_e\frac{1}{(z+c)^2}\right]_{x=0}. \tag{1·12}$$

Boundary values of these stresses being obtained in the above forms their values may be obtained at any other point by putting in (1·5)

$$f_1(z) = \frac{Q}{2\pi}\frac{i}{(z+c)^2}, \quad f_2(z) = \frac{Q}{2\pi}\frac{1}{(z+c)^2}. \tag{1·13}$$

Thus

$$\left.\begin{aligned}\widehat{xx}_2 &= \frac{Q}{2\pi} R_e\left[\frac{4ix}{(z+c)^3}+\frac{i}{(z+c)^2}\right],\\ \widehat{xy}_2 &= \frac{Q}{2\pi} R_e\left[-\frac{4x}{(z+c)^3}+\frac{1}{(z+c)^2}\right],\\ \widehat{yy}_2 &= \frac{Q}{2\pi} R_e\left[\frac{3i}{(z+c)^2}-\frac{4ix}{(z+c)^3}\right].\end{aligned}\right\} \tag{1·14}$$

Combining the results given in (1·9) and (1·14), then

$$\left.\begin{aligned}\widehat{xx} &= \widehat{xx}_1+\widehat{xx}_2\\ &= \frac{Q}{\pi}\left[\frac{(x-c)\,y}{r_1^4}+\frac{(x+c)\,y}{r_2^4}+\frac{2xy\{3(x+c)^2-y^2\}}{r_2^6}\right],\\ \widehat{xy} &= \widehat{xy}_1+\widehat{xy}_2\\ &= \frac{Q}{2\pi}\left[\frac{y^2-(x-c)^2}{r_1^4}+\frac{(x+c)^2-y^2}{r_2^4}-\frac{4x(x+c)\,\{(x+c)^2-3y^2\}}{r_2^6}\right],\\ \widehat{yy} &= \widehat{yy}_1+\widehat{yy}_2\\ &= -\frac{Q}{\pi}\left[\frac{(x-c)\,y}{r_1^4}-\frac{3y(x+c)}{r_2^4}+\frac{2xy\{3(x+c)^2-y^2\}}{r_2^6}\right].\end{aligned}\right\} \tag{1·15}$$

(*c*) *Stresses due to a point of pressure situated inside a semi-infinite plate*

Let there be a point of pressure at A $(c, 0)$ so that A being in a circular cavity there is uniform pressure of amount p_0/r^2 over the cavity (Love 1927). If the plate had extended indefinitely in all directions there would be, in consequence of this point of pressure, the stresses

$$\widehat{xx}_1 = p_0\frac{y^2-(x-c)^2}{r_1^4},\quad \widehat{xy}_1 = -\frac{2p_0(x-c)\,y}{r_1^4},\quad \widehat{yy}_1 = p_0\frac{(x-c)^2-y^2}{r_1^4}. \tag{1·16}$$

On the line $x = 0$ it follows that

$$[\widehat{xx}_1]_{x=0} = p_0\frac{y^2-c^2}{(c^2+y^2)^2},\quad [\widehat{xy}_1]_{x=0} = 2p_0\frac{cy}{(c^2+y^2)^2}. \tag{1·17}$$

The superposed stresses $\widehat{xx}_2$, $\widehat{xy}_2$ that will make the edge $x = 0$ free from stresses would be such that

$$\left.\begin{aligned}[\widehat{xx}_2]_{x=0} &= p_0\frac{(c^2-y^2)}{(c^2+y^2)^2} = \left[R_e\frac{p_0}{(z+c)^2}\right]_{x=0},\\ [\widehat{xy}_2]_{x=0} &= -\frac{2p_0cy}{(c^2+y^2)^2} = \left[R_e-\frac{p_0 i}{(z+c)^2}\right]_{x=0}.\end{aligned}\right\} \tag{1·18}$$

The complex functions in the above expressions are so chosen that they have singularities at the point $(-c, 0)$. To obtain the stresses $\widehat{xx}_2$, $\widehat{xy}_2$, $\widehat{yy}_2$ it is necessary to put in (1·5)

$$f_1(z) = \frac{p_0}{(z+c)^2},\quad f_2(z) = -\frac{p_0 i}{(z+c)^2}. \tag{1·19}$$

Thus

$$\left.\begin{aligned} \widehat{xx}_2 &= p_0 R_e\left[\frac{4x}{(z+c)^3}+\frac{1}{(z+c)^2}\right], \\ \widehat{xy}_2 &= p_0 R_e\left[\frac{4xi}{(z+c)^3}-\frac{i}{(z+c)^2}\right], \\ \widehat{yy}_2 &= p_0 R_e\left[\frac{3}{(z+c)^2}-\frac{4x}{(z+c)^3}\right]. \end{aligned}\right\} \qquad (1\cdot20)$$

Hence combining (1·20) with (1·16)

$$\left.\begin{aligned} \widehat{xx} &= \widehat{xx}_1+\widehat{xx}_2 \\ &= p_0\left[\frac{y^2-(x-c)^2}{r_1^4}+\frac{(x+c)^2-y^2}{r_2^4}+\frac{4x(x+c)\{(x+c)^2-3y^2\}}{r_2^6}\right], \\ \widehat{xy} &= \widehat{xy}_1+\widehat{xy}_2 \\ &= -2p_0\left[\frac{(x-c)\,y}{r_1^4}+\frac{(x+c)\,y}{r_2^4}+\frac{2xy\{y^2-(x+c)^2\}}{r_2^6}\right], \\ \widehat{yy} &= \widehat{yy}_1+\widehat{yy}_2 \\ &= p_0\left[\frac{(x-c)^2-y^2}{r_1^4}+\frac{3\{(x+c)^2-y^2\}}{r_2^4}-\frac{4x(x+c)\{(x+c)^2-3y^2\}}{r_2^6}\right]. \end{aligned}\right\} \qquad (1\cdot21)$$

On the edge $x = 0$

$$[\widehat{yy}]_{x=0} = \frac{4p_0(c^2-y^2)}{(c^2+y^2)^2},$$

which shows that on this edge $\widehat{yy} = 0$ when $y = \pm c$; and $\widehat{yy}$ is negative when $|y| > c$.

SECTION II

(*a*) *Stresses due to an isolated couple acting at a point inside a thin semi-infinite plate of certain non-isotropic material*

In the paper II referred to above, problems of a plate having a particular kind of aelotropy were considered. In this type it was assumed that the Poisson's ratios σ_1 and σ_2, and Young's moduli E_1 and E_2 corresponding to two perpendicular directions taken as the axes of x and y, are connected by the relations

$$\frac{\sigma_1}{\sigma_2} = \frac{E_2}{E_1} = k^2 \quad \text{(say).} \qquad (2\cdot1)$$

In a semi-infinite plate of such material bounded by the axis of x it is found that (see paper II, art. X)

$$\left.\begin{aligned} \widehat{yy} &= \frac{1}{k-1}R_e[kf(z_1)-f(z)+k\{\psi(z_1)-\psi(z)\}], \\ \widehat{xy} &= \frac{1}{k-1}R_e[-i\{f(z_1)-f(z)\}+i\{k\psi(z)-\psi(z_1)\}], \\ \widehat{xx} &= \frac{1}{k-1}R_e\left[f(z)-\frac{1}{k}f(z_1)+k\psi(z)-\frac{1}{k}\psi(z_1)\right], \end{aligned}\right\} \qquad (2\cdot2)$$

where

$$z = x+iy, \quad z_1 = x+i\frac{y}{k}, \quad [\widehat{yy}]_{y=0} = [R_e f(z)]_{y=0}, \quad [\widehat{xy}]_{y=0} = [R_e i\psi(z)]_{y=0}. \qquad (2\cdot3)$$

For a couple of moment Q acting at the point $(0, c)$ in a thin plate of such a material as described above which extends indefinitely in all directions, the following corresponding stresses are found:

$$\widehat{yy}_1 = \frac{Q}{\pi}\frac{x(y-c)}{[x^2+(y-c)^2]^2}, \quad \widehat{xy}_1 = \frac{Q}{2\pi}\frac{x^2-(y-c)^2}{[x^2+(y-c)^2]^2}, \quad \widehat{xx}_1 = -\frac{Q}{\pi}\frac{x(y-c)}{[x^2+(y-c)^2]^2}. \tag{2·4}$$

On the edge $y = 0$ are found the stresses

$$[\widehat{yy}_1]_{y=0} = -\frac{Q}{\pi}\frac{xc}{(x^2+c^2)^2}, \quad [\widehat{xy}_1]_{y=0} = \frac{Q}{2\pi}\frac{x^2-c^2}{(x^2+c^2)^2}. \tag{2·5}$$

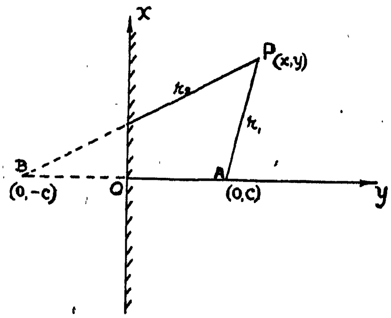


Figure 2

To counteract the above stresses on the boundary it will be necessary to add a stress system of which the components $\widehat{yy}_2$, $\widehat{xy}_2$ should be such that

$$[\widehat{yy}_2]_{y=0} = \frac{Q}{\pi}\frac{xc}{(x^2+c^2)^2}, \quad [\widehat{xy}_2]_{y=0} = -\frac{Q}{2\pi}\frac{x^2-c^2}{(x^2+c^2)^2}. \tag{2·6}$$

It being assumed that the superposed stress system has no singularities at any point inside the plate, then

$$[\widehat{yy}_2]_{y=0} = \frac{Q}{2\pi}\left[R_e\frac{i}{(z+ic)^2}\right]_{y=0}, \quad [\widehat{xy}_2]_{y=0} = -\frac{Q}{2\pi}\left[R_e\frac{1}{(z+ic)^2}\right]_{y=0}. \tag{2·7}$$

Hence put
$$f(z) = \psi(z) = \frac{Q}{2\pi}\frac{i}{(z+ic)^2}$$
in (2·2) and obtain

$$\left.\begin{aligned} \widehat{yy}_2 &= \frac{Q}{2\pi(k-1)}R_e\left[\frac{2ki}{(z_1+ic)^2}-\frac{(k+1)i}{(z+ic)^2}\right], \\ \widehat{xy}_2 &= \frac{Q}{2\pi(k-1)}R_e\left[\frac{2}{(z_1+ic)^2}-\frac{k+1}{(z+ic)^2}\right], \\ \widehat{xx}_2 &= \frac{Q}{2\pi(k-1)}R_e\left[\frac{(k+1)i}{(z+ic)^2}-\frac{2i}{k(z_1+ic)^2}\right]. \end{aligned}\right\} \tag{2·8}$$

Putting now $r_1^2 = x^2+(y-c)^2$, $r_2^2 = x^2+(y+c)^2$, $r_k^2 = x^2+\left(\frac{y}{k}+c\right)^2$, and combining (2·4) with (2·8) it is seen that

$$\left.\begin{aligned}
\widehat{yy} &= \widehat{yy}_1+\widehat{yy}_2\\
&= \frac{Q}{\pi}\left[\frac{x(y-c)}{r_1^4}+\frac{2x(y+kc)}{(k-1)\,r_k^4}-\frac{(k+1)}{(k-1)}\frac{x(y+c)}{r_2^4}\right],\\
\widehat{xy} &= \widehat{xy}_1+\widehat{xy}_2\\
&= \frac{Q}{2\pi}\left[\frac{x^2-(y-c)^2}{r_1^4}+\frac{2\{x^2-(y/k+c)^2\}}{(k-1)\,r_k^4}-\frac{(k+1)}{(k-1)}\frac{\{x^2-(y+c)^2\}}{r_2^4}\right],\\
\widehat{xx} &= \widehat{xx}_1+\widehat{xx}_2\\
&= -\frac{Q}{\pi}\left[\frac{x(y-c)}{r_1^4}-\frac{(k+1)}{(k-1)}\frac{x(y+c)}{r_2^4}+\frac{2x}{k^2}\frac{(y+kc)}{(k-1)\,r_k^4}\right].
\end{aligned}\right\} \tag{2·9}$$

On the bounding edge $y=0$

$$[\widehat{xx}]_{y=0}=\frac{2Q(k+1)}{k\pi}\frac{cx}{(c^2+x^2)^2}, \tag{2·10}$$

which shows that the numerically greatest values of $\widehat{xx}$ occur at the points given by $x=\pm\frac{c}{\sqrt{3}}$, these values being given by

$$\widehat{xx}=\pm\frac{3\sqrt{3}\,(k+1)\,Q}{8\pi kc^2}. \tag{2·11}$$

(*b*) *Stresses due to a point of pressure situated inside a semi-infinite plate of the type described above*

Let there be a point of pressure at the point $(0,c)$. If the plate had extended indefinitely in all directions, the stresses due to this singularity would have been (Love 1927)

$$\widehat{yy}_1=p_0\frac{x^2-(y-c)^2}{r_1^4},\quad \widehat{xy}_1=-2p_0\frac{x(y-c)}{r_1^4},\quad \widehat{xx}_1=p_0\frac{(y-c)^2-x^2}{r_1^4}, \tag{2·12}$$

where $r_1^2=x^2+(y-c)^2$. Then, on the edge $y=0$,

$$[\widehat{yy}_1]_{y=0}=p_0\frac{x^2-c^2}{(x^2+c^2)^2},\quad [\widehat{xy}_1]_{y=0}=2p_0\frac{xc}{(x^2+c^2)^2}. \tag{2·13}$$

In order that this edge may be free from stress, the stress components $\widehat{yy}_2$ and $\widehat{xy}_2$ of the superposed system should be such that

$$[\widehat{yy}_2]_{y=0}=-p_0\frac{x^2-c^2}{(x^2+c^2)^2},\quad [\widehat{xy}_2]_{y=0}=-2p_0\frac{xc}{(x^2+c^2)^2}. \tag{2·14}$$

These expressions may be written in the forms

$$[\widehat{yy}_2]_{y=0}=-p_0\left[R_e\frac{1}{(z+ic)^2}\right]_{y=0},\quad [\widehat{xy}_2]_{y=0}=-p_0\left[R_e\frac{i}{(z+ic)^2}\right]_{y=0}. \tag{2·15}$$

Hence putting
$$f(z) = \psi(z) = -\frac{p_0}{(z+ic)^2} \tag{2·16}$$
in (2·2) then
$$\left.\begin{aligned}
\widehat{yy}_2 &= \frac{p_0}{k-1}R_e\left[-\frac{2k}{(z_1+ic)^2}+\frac{k+1}{(z+ic)^2}\right],\\
\widehat{xy}_2 &= \frac{p_0}{k-1}R_e\left[\frac{2i}{(z_1+ic)^2}-\frac{(k+1)i}{(z+ic)^2}\right],\\
\widehat{xx}_2 &= \frac{p_0}{k-1}R_e\left[\frac{2}{k(z_1+ic)^2}-\frac{k+1}{(z+ic)^2}\right].
\end{aligned}\right\} \tag{2·17}$$
Combining (2·12) with (2·17) then
$$\left.\begin{aligned}
\widehat{yy} &= \widehat{yy}_1+\widehat{yy}_2\\
&= p_0\left[\frac{x^2-(y-c)^2}{r_1^4}+\frac{k+1}{k-1}\frac{\{x^2-(y+c)^2\}}{r_2^4}-\frac{2k}{k-1}\frac{\{x^2-(y/k+c)^2\}}{r_k^4}\right],\\
\widehat{xy} &= \widehat{xy}_1+\widehat{xy}_2\\
&= p_0\left[\frac{-2x(y-c)}{r_1^4}-\frac{(k+1)}{(k-1)}\frac{2x(y+c)}{r_2^4}+\frac{4x}{k-1}\frac{(y/k+c)}{r_k^4}\right],\\
\widehat{xx} &= \widehat{xx}_1+\widehat{xx}_2\\
&= p_0\left[\frac{(y-c)^2-x^2}{r_1^4}+\frac{2\{x^2-(y/k+c)^2\}}{k(k-1)\,r_k^4}-\frac{(k+1)}{(k-1)}\frac{\{x^2-(y+c)^2\}}{r_2^4}\right].
\end{aligned}\right\} \tag{2·18}$$
Hence, on the edge $y = 0$,
$$[\widehat{xx}]_{y=0} = 2p_0\frac{(k+1)}{k}\frac{c^2-x^2}{(c^2+x^2)^2},$$
which shows that on this line
$$\widehat{xx} = 0, \quad \text{when} \quad x = \pm c.$$

SECTION III

Stresses in a circular plate of isotropic material

It has been shown in paper I [art. 2 (*a*)] that the equations of equilibrium can be put in the forms
$$\nabla_1^2[r\,.\,\widehat{rx}]+\frac{\partial}{\partial x}\left[r\frac{\partial\Theta}{\partial r}-\Theta\right] = 0, \quad \nabla_1^2[r\,.\,\widehat{ry}]+\frac{\partial}{\partial y}\left[r\frac{\partial\Theta}{\partial r}-\Theta\right] = 0, \tag{3·1}$$
where ∇_1^2 stands for $\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}$, $\widehat{rx}$ and $\widehat{ry}$ are respectively the x component and the y component of the stresses at the point (r, θ) on the circle of radius r and $\Theta = \widehat{xx}+\widehat{yy}$. Since Θ is a plane harmonic function then
$$\Theta = R_e f(z), \tag{3·2}$$
z standing for $x+iy$.

Substituting this value of Θ in (3·1), then
$$\nabla_1^2[r\,.\,\widehat{rx}] = R_e\frac{\partial}{\partial z}[f(z)-zf'(z)], \quad \nabla_1^2[r\,.\,\widehat{ry}] = R_e i\frac{\partial}{\partial z}[f(z)-zf'(z)]. \tag{3·3}$$

To find the solutions of the above equations suitable for circular boundaries put

$$r\,.\,\widehat{rx} = R_e\left[\frac{r^2-a^2}{4}\phi(z)+aL(z)\right], \quad r\,.\,\widehat{ry} = R_e\left[\frac{r^2-a^2}{4}i\psi(z)+aM(z)\right]. \tag{3·4}$$

where $\phi(z)$, $\psi(z)$, $L(z)$, $M(z)$ are functions of z.

The functions $L(z)$ and $M(z)$ are such that

$$[\widehat{rx}]_{r=a} = [R_e L(z)]_{r=a}, \quad [\widehat{ry}]_{r=a} = [R_e M(z)]_{r=a}. \tag{3·5}$$

On substituting the expressions (3·4) in (3·3) it is found that these equations are satisfied if

$$z\phi(z) = z\psi(z) = f(z)-zf'(z).$$

Hence as the solutions of (3·3) one can write

$$\left.\begin{aligned} r\,.\,\widehat{rx} &= R_e\left[\frac{r^2-a^2}{4}\left\{\frac{f(z)-zf'(z)}{z}\right\}+aL(z)\right],\\ r\,.\,\widehat{ry} &= R_e\left[\frac{r^2-a^2}{4}i\left\{\frac{f(z)-zf'(z)}{z}\right\}+aM(z)\right].\end{aligned}\right\} \tag{3·6}$$

Since $\qquad \Theta = \dfrac{\partial}{\partial x}[r\,.\,\widehat{rx}]+\dfrac{\partial}{\partial y}[r\,.\,\widehat{ry}].$

then $\qquad R_e[f(z)+zf'(x)] = 2aR_e[L'(z)+iM'(z)],$

which gives

$$f(z) = 2a\frac{L(z)+iM(z)}{z} \tag{3·7}$$

except in the case* $\quad f(z)+zf'(z) = 0, \quad$ or when $\quad f(z)\propto\dfrac{1}{z}. \qquad$ (3·8)

Thus, in general, $L(z)$ and $M(z)$ being known from the boundary values of $\widehat{rx}$ and $\widehat{ry}$, $f(z)$ can be found from (3·7) and hence the stresses at every point of the plate can be determined with the help of (3·6).

(*a*) *Stresses due to couples acting in the interior of a circular plate*

Let a couple of moment Q act at the point $(c, 0)$ within a circular plate of radius a. For equilibrium let there be another couple of equal and opposite moment operating at the centre,† the axes of the couples being perpendicular to the plane of the plate. The stresses due to these couples operating on a plate extended indefinitely in all directions are given by

$$\left.\begin{aligned} r\,.\,\widehat{rx}_1 &= x\,.\,\widehat{xx}_1+y\,.\,\widehat{xy}_1 = \frac{Qy}{2\pi}\left[\frac{r^2-c^2}{r_1^4}-\frac{1}{r^2}\right],\\ r\,.\,\widehat{ry}_1 &= x\,.\,\widehat{xy}_1+y\,.\,\widehat{yy}_1 = \frac{Q}{2\pi}\left[\frac{-x(r^2+c^2)+2cr^2}{r_1^4}+\frac{x}{r^2}\right],\end{aligned}\right\} \tag{3·9}$$

where $\qquad r_1^2 = (x-c)^2+y^2 \quad$ and $\quad r^2 = x^2+y^2.$

---

* This case can be treated separately (see Paper I, art. 3 (*b*)).

† The solution can also be obtained for a different position of the equilibrating couple.

Let A be the point $(c, 0)$ and B its inverse with respect to the circle of radius a (figure 3). Then the co-ordinates of B are $(a^2/c, 0)$. If P be any point (x, y) on the circular boundary, the above stresses at P will be

$$\left.\begin{aligned}[r\,.\,\widehat{rx}_1]_{r=a} &= \frac{Q}{2\pi}\left[\frac{y(a^2-c^2)}{AP^4}-\frac{y}{a^2}\right] = \frac{Q}{2\pi}\left[\frac{a^4(a^2-c^2)\,y}{c^4BP^4}-\frac{y}{a^2}\right],\\ [r\,.\,\widehat{ry}_1]_{r=a} &= \frac{Q}{2\pi}\left[\frac{-x(a^2+c^2)+2ca^2}{AP^4}+\frac{x}{a^2}\right] = \frac{Q}{2\pi}\left[\frac{a^4}{c^4}\frac{\{-x(a^2+c^2)+2ca^2\}}{BP^4}+\frac{x}{a^2}\right].\end{aligned}\right\} \tag{3·10}$$

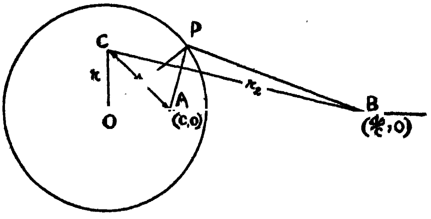


FIGURE 3

To make the circular boundary free the stresses $\widehat{rx}_2$, $\widehat{ry}_2$ may be added which have singularities at the point $(a^2/c, 0)$ and of which the boundary values are equal and opposite to those given in (3·10). Hence

$$\left.\begin{aligned}[r\,.\,\widehat{rx}_2]_{r=a} &= \frac{Q}{2\pi}\left[\frac{a^4(c^2-a^2)\,y}{c^4BP^4}+\frac{y}{a^2}\right],\\ &= \frac{Q}{2\pi}\left[R_e\left\{\frac{a^2}{c^2}\frac{iz}{(z-a^2/c)^2}-\frac{iz}{a^2}\right\}\right]_{r=a},\\ [r\,.\,\widehat{ry}_2]_{r=a} &= \frac{Q}{2\pi}\left[\frac{a^4}{c^4}\frac{\{x(a^2+c^2)-2ca^2\}}{BP^4}-\frac{x}{a^2}\right],\\ &= \frac{Q}{2\pi}\left[R_e\left\{\frac{a^2}{c^2}\frac{z}{(z-a^2/c)^2}-\frac{z}{a^2}\right\}\right]_{r=a}.\end{aligned}\right\} \tag{3·11}$$

To find the stresses $\widehat{rx}_2$, $\widehat{ry}_2$ from the results (3·6) it will be necessary to write

$$aL(z) = \frac{Q}{2\pi}\left[\frac{a^2}{c^2}\frac{iz}{(z-a^2/c)^2}-\frac{iz}{a^2}\right], \quad aM(z) = \frac{Q}{2\pi}\left[\frac{a^2}{c^2}\frac{z}{(z-a^2/c)^2}-\frac{z}{a^2}\right]. \tag{3·12}$$

Putting these values in (3·7), then

$$f(z) = \frac{2Q}{\pi}\left[\frac{a^2}{c^2}\frac{i}{(z-a^2/c)^2}-\frac{i}{a^2}\right]. \tag{3·13}$$

This value of $f(z)$ being substituted in (3·6) then

$$\left.\begin{aligned}r\,.\,\widehat{rx}_2 &= \frac{Q}{2\pi}\frac{a^2}{c^2}R_e\left[(r^2-a^2)\,i\left\{-\frac{c^2}{a^4(z-a^2/c)}+\frac{c}{a^2(z-a^2/c)^2}+\frac{2}{(z-a^2/c)^3}\right\}+\frac{iz}{(z-a^2/c)^2}-\frac{izc^2}{a^4}\right],\\ r\,.\,\widehat{ry}_2 &= \frac{Q}{2\pi}\frac{a^2}{c^2}R_e\left[-(r^2-a^2)\left\{-\frac{c^2}{a^4(z-a^2/c)}+\frac{c}{a^2(z-a^2/c)^2}+\frac{2}{(z-a^2/c)^3}\right\}+\frac{z}{(z-a^2/c)^2}-\frac{c^2z}{a^4}\right].\end{aligned}\right\} \tag{3·14}$$

If one writes $$r_2^2 = (x-a^2/c)^2+y^2,$$
and combines (3·10) with (3·14), then

$$\left.\begin{aligned}
r\,.\,\widehat{rx} &= r\,.\,\widehat{rx}_1+r\,.\,\widehat{rx}_2\\
&= \frac{Qy}{2\pi}\left[\frac{r^2-c^2}{r_1^4}-\frac{1}{r^2}+(r^2-a^2)\left\{-\frac{1}{a^2r_2^2}+\frac{2(x-a^2/c)}{cr_2^4}+\frac{2a^2}{c^2}\frac{[3(x-a^2/c)^2-y^2]}{r_2^6}\right\}\right.\\
&\qquad\left.+\frac{a^2}{c^2}\frac{(r^2-a^4/c^2)}{r_2^4}+\frac{1}{a^2}\right],\\
r\,.\,\widehat{ry} &= r\,.\,\widehat{ry}_1+r\,.\,\widehat{ry}_2\\
&= \frac{Q}{2\pi}\left[\frac{2cr^2-x(r^2+c^2)}{r_1^4}+\frac{x}{r^2}+(a^2-r^2)\left\{-\frac{x-a^2/c}{a^2r_2^2}+\frac{(x-a^2/c)^2-y^2}{cr_2^4}\right.\right.\\
&\qquad\left.\left.+\frac{2a^2(x-a^2/c)\,[(x-a^2/c)^2-3y^2]}{c^2r_2^6}\right\}+\frac{a^2}{c_2}\frac{\{x(r^2+a^4/c^2)-(2a^2r^2/c)}{r_2^4}-\frac{x}{a^2}\right].
\end{aligned}\right\}\qquad(3{\cdot}15)$$

The hoop stress $\widehat{\theta\theta}$ at any point on the boundary can be found out from the boundary value of Θ $(=\widehat{rr}+\widehat{\theta\theta})$. The contribution to Θ by $\widehat{xx}_1$ and $\widehat{yy}_1$ is nil. Since at the circular edge $\widehat{rr}=0$, then
$$[\widehat{\theta\theta}]_{r=a} = [\Theta]_{r=a} = [R_e f(z)]_{r=a},$$
which in the present case is equal to
$$\left\{R_e\left[\frac{2Q}{\pi}\frac{a^2}{c^2}\frac{i}{(z-a^2/c)^2}\right]\right\}_{r=a} = -\frac{4Q}{\pi}c\frac{(a-c\cos\theta)\sin\theta}{(a^2+c^2-2ac\cos\theta)^2},\qquad(3{\cdot}16)$$
where (a,θ) are the co-ordinates of the point on the boundary. The above expression evidently vanishes when $\theta=0$ or π.

(*b*) *Stresses due to forces acting in the interior of a circular plate*

Let an isolated force T act at the point A $(c,0)$ in the positive direction of the x-axis (figure 3), while an equal and opposite force acting at the origin balances it. If the plate has extended indefinitely in all directions the corresponding stress components would follow:

$$\left.\begin{aligned}
\widehat{xx}_1 &= \frac{T}{4\pi}\left[-\frac{(3+\sigma)(x-c)}{r_1^2}+\frac{2(1+\sigma)(x-c)y^2}{r_1^4}+\frac{(3+\sigma)x}{r^2}-\frac{2(1+\sigma)xy^2}{r^4}\right],\\
\widehat{xy}_1 &= -\frac{T}{4\pi}\left[\frac{(3+\sigma)y}{r_1^2}-\frac{2(1+\sigma)y^3}{r_1^4}-\frac{(3+\sigma)y}{r^2}+\frac{2(1+\sigma)y^3}{r^4}\right],\\
\widehat{yy}_1 &= \frac{T}{4\pi}\left[\frac{(1-\sigma)(x-c)}{r_1^2}-\frac{2(1+\sigma)y^2(x-c)}{r_1^4}-\frac{(1-\sigma)x}{r^2}+\frac{2(1+\sigma)xy^2}{r^4}\right].
\end{aligned}\right\}\qquad(3{\cdot}17)$$

These results give

$$\left.\begin{aligned}
r\,.\,\widehat{rx}_1 &= \frac{T}{4\pi}\left[\frac{(3+\sigma)(x-r^2/c)c}{r_1^2}+(3+\sigma)+\frac{(1+\sigma)y^2}{r_1^2}+\frac{(1+\sigma)y^2(r^2-c^2)}{r_1^4}-\frac{2(1+\sigma)y^2}{r^2}\right],\\
r\,.\,\widehat{ry}_1 &= -\frac{T}{4\pi}\left[\frac{(1-\sigma)cy}{r_1^2}+\frac{2(1+\sigma)xy}{r_1^2}-\frac{2(1+\sigma)cy^3}{r_1^4}-\frac{2(1+\sigma)xy}{r^2}\right].
\end{aligned}\right\}\qquad(3{\cdot}18)$$

Then, at a point $P(x,y)$ on the circular edge $r = a$,

$$\left.\begin{aligned}
[r\,.\,\widehat{rx}_1]_{r=a} &= \frac{T}{4\pi}\left[\frac{(3+\sigma)(x-a^2/c)\,c}{AP^2}+(3+\sigma)+\frac{(1+\sigma)\,y^2}{AP^2}+\frac{(1+\sigma)\,y^2\,(a^2-c^2)}{AP^4}-\frac{2(1+\sigma)\,y^2}{a^2}\right],\\
[r\,.\,\widehat{ry}_1]_{r=a} &= -\frac{T}{4\pi}\left[\frac{(1-\sigma)\,cy}{AP^2}+\frac{2(1+\sigma)\,xy}{AP^2}-\frac{2(1+\sigma)\,cy^3}{AP^4}-\frac{2(1+\sigma)\,xy}{a^2}\right].
\end{aligned}\right\} \tag{3·19}$$

The stress components $\widehat{rx}_2$, $\widehat{ry}_2$ to be added for nullifying these stresses on the boundary should be such that

$$\begin{aligned}
[r\,.\,\widehat{rx}_2]_{r=a} &= -\frac{T}{4\pi}\left[\frac{(3+\sigma)\,a^2(x-a^2/c)}{cBP^2}+(3+\sigma)+\frac{(1+\sigma)\,y^2a^2}{c^2BP^2}+\frac{(1+\sigma)\,a^4}{c^4}\,\frac{y^2(a^2-c^2)}{BP^4}-\frac{2(1+\sigma)\,y^2}{a^2}\right]\\
[r\,.\,\widehat{ry}_2]_{r=a} &= \frac{T}{4\pi}\left[\frac{(1-\sigma)\,a^2y}{cBP^2}+\frac{2(1+\sigma)\,a^2xy}{c^2BP^2}-\frac{2(1+\sigma)\,a^4y^3}{c^3BP^4}-\frac{2(1+\sigma)\,xy}{a^2}\right].
\end{aligned} \tag{3·20}$$

The above results can then be put in the forms

$$\left.\begin{aligned}
[r\,.\,\widehat{rx}_2]_{r=a} &= -\frac{T}{4\pi}\left[R_e\left\{\frac{(3+\sigma)\,a^2}{c(z-a^2/c)}+(3+\sigma)+\frac{(1+\sigma)}{2c}\right.\right.\\
&\qquad\left.\left.\times\frac{z^2-a^2}{(z-a^2/c)}-\frac{(1+\sigma)\,a^2(z^2-a^2)}{2c^2(z-a^2/c)^2}+\frac{(1+\sigma)}{a^2}(z^2-a^2)\right\}\right]_{r=a},\\
&= -\frac{T}{4\pi}\left[R_e\left\{\frac{(3+\sigma)\,z}{z-a^2/c}+\frac{(1+\sigma)\,z(z^2-a^2)}{a^2(z-a^2/c)}-\frac{(1+\sigma)\,z(z^2-a^2)}{2c(z-a^2/c)^2}\right\}\right]_{r=a}\\
&= -\frac{T}{4\pi}\left[R_e\left\{\frac{2z}{z-a^2/c}+\frac{(1+\sigma)\,z^3}{a^2(z-a^2/c)}-\frac{(1+\sigma)\,z(z^2-a^2)}{2c(z-a^2/c)^2}\right\}\right]_{r=a},\\
[r\,.\,\widehat{ry}_2]_{r=a} &= -\frac{T}{4\pi}\left[R_e\left\{-\frac{(1-\sigma)\,iz}{z-a^2/c}-\frac{(1+\sigma)\,iz^2}{c(z-a^2/c)}+\frac{(1+\sigma)\,iz(z^2-a^2)}{2c(z-a^2/c)^2}-\frac{(1+\sigma)\,iz^2}{a^2}\right\}\right]_{r=a}\\
&= -\frac{T}{4\pi}\left[R_e\left\{-\frac{(1-\sigma)\,iz}{z-a^2/c}-\frac{(1+\sigma)\,iz^3}{a^2(z-a^2/c)}+\frac{(1+\sigma)\,iz(z^2-a^2)}{2c(z-a^2/c)^2}\right\}\right]_{r=a}.
\end{aligned}\right\} \tag{3·21}$$

Putting now in (3·7)

$$\left.\begin{aligned}
aL(z) &= -\frac{T}{4\pi}\left[\frac{2z}{z-a^2/c}+\frac{(1+\sigma)\,z^3}{a^2(z-a^2/c)}-\frac{(1+\sigma)\,z(z^2-a^2)}{2c(z-a^2/c)^2}+\frac{c}{2a^2}(1-\sigma)\frac{z^2-a^2}{z}\right],\\
aM(z) &= -\frac{T}{4\pi}\left[-\frac{(1-\sigma)\,iz}{z-a^2/c}-\frac{(1+\sigma)\,iz^3}{a^2(z-a^2/c)}+\frac{(1+\sigma)\,iz(z^2-a^2)}{2c(z-a^2/c)^2}-\frac{ic}{2a^2}(1-\sigma)\frac{z^2+a^2}{z}\right],
\end{aligned}\right\} \tag{3·22}$$

it follows that

$$\begin{aligned}
f(z) &= 2a\,\frac{L(z)+iM(z)}{z}\\
&= -\frac{T}{2\pi}\left[\frac{(3-\sigma)}{z-a^2/c}+\frac{2(1+\sigma)\,z^2}{a^2(z-a^2/c)}-\frac{(1+\sigma)\,(z^2-a^2)}{c(z-a^2/c)^2}+\frac{c}{a^2}(1-\sigma)\right]\\
&= -\frac{T}{2\pi}\left[\frac{2(1-\sigma)}{z-a^2/c}+\frac{2(1+\sigma)\,z^2}{a^2(z-a^2/c)}-\frac{(1+\sigma)}{c}\,\frac{z(z-c)}{(z-a^2/c)^2}+\frac{c}{a^2}(1-\sigma)\right].
\end{aligned} \tag{3·23}$$

Substituting the values (3·22) and (3·23) in (3·6) the values of the components $r.\widehat{rx}_2$ and $r.\widehat{ry}_2$ are obtained, which, combined with the components $r.\widehat{rx}_1$ and $r.\widehat{ry}_1$, give us the required stress distribution in the plate.

From the expressions (3·17) and (3·23) then

$$\begin{aligned}\Theta &= (\widehat{xx}_1+\widehat{xx}_2)+(\widehat{yy}_1+\widehat{yy}_2) = \widehat{xx}_1+\widehat{yy}_1+R_e[f(z)] \\ &= -\frac{T}{2\pi}\left[\frac{(1+\sigma)(x-c)}{r_1^2}-\frac{(1+\sigma)x}{r^2}+\frac{2(1-\sigma)(x-a^2/c)}{r_2^2}+\frac{2(1+\sigma)}{a^2r_2^2}\{(x^2-y^2)(x-a^2/c)+2xy^2\}\right. \\ &\qquad \left.+\frac{c}{a^2}(1-\sigma)-\frac{(1+\sigma)}{cr_2^4}\{(x^2-cx-y^2)[(x-a^2/c)^2-y^2]+2y^2(x-a^2/c)(2x-c)\}\right]. \quad (3{\cdot}24)\end{aligned}$$

Since $\Theta = \widehat{rr}+\widehat{\theta\theta}$ and $(\widehat{rr})_{r=a} = 0$ it follows that at any point (x, y) on the boundary

$$\begin{aligned}[\widehat{\theta\theta}]_{r=a} &= [\Theta]_{r=a} \\ &= -\frac{T}{2\pi}\left[\frac{(1+\sigma)(x-c)}{AP^2}-\frac{(1+\sigma)x}{a^2}+\frac{2(1-\sigma)(x-a^2/c)}{BP^2}\right. \\ &\qquad \left.+\frac{c}{a^2}(1-\sigma)-\frac{3(1+\sigma)(x^2-cx-y^2)}{cBP^2}+\frac{4(1+\sigma)a^2y^2(x-c)}{c^2BP^4}\right]. \quad (3{\cdot}25)\end{aligned}$$

REFERENCES

Ghosh, S. 1937 *Bull. Calcutta Math. Soc.* **29**, 177.
Love, A. E. H. 1927 *Mathematical Theory of Elasticity*, Art. 152.
Melan, E. 1932 *Z. Angew. Math. Mech.* **12**, 343.
Sen, B. 1938 *Phil. Mag.* **26**, 98–119.
Sen, B. 1939 *Phil. Mag.* **27**, 596–604.
Timoshenko, S. 1932 *Theory of Elasticity*, Art. 34.

The mechanism of ionospheric ionization

By R. v. d. R. Woolley

(*Communicated by S. Chapman, F.R.S.—Received* 10 *October* 1945)

The available mechanisms for the production of electrons in the three regions of the ionosphere are discussed with special reference to the question whether it is possible to account for the observed electron densities without supposing that the sun emits far more energy in the remote ultra-violet spectrum than would be emitted by a black body at 6000°. The contributions to electron densities made by metastable states of atoms and molecules are examined. It is concluded that the observed electron densities may be accounted for without requiring high solar energy in the ultra-violet if the effective recombination coefficient in the F_2 region is 10^{-11}. The F_2 region is attributed to the ionization of atomic oxygen, and the E region to the ionization of molecular oxygen. The electrons forming the F_1 region are supposed to be provided by metastable N_2 or by NO.

1. The electron density in the earth's atmosphere considered as a function of height exhibits several maxima which are observed by the reflexion of radio waves sent out from ground level. Broadly speaking three such maxima are observed: the neighbourhood of each maximum is called a region, and the regions are named E, F_1 and F_2 in order of increasing height. The E region reflects all waves having frequencies less than the critical frequency of the E region. Waves above this frequency and below the critical frequency of the F_1 region are reflected by the F_1 region, and waves of still greater frequency are reflected by the F_2 region. Waves of frequency greater than the critical frequency of the F_2 region are not reflected at all. Since the critical frequency is proportional to the square root of the maximum electron density in the layer, the maximum electron densities increase from E to F_2. It follows that there may well be unobservable maxima in the electron density.

Suppose that for a particular absorption spectrum in the atmosphere capable of ionizing an atom or molecule the absorption coefficient $k(\nu)$ is independent of ν for all frequencies greater than ν_0 and is zero for $\nu < \nu_0$. The number of electrons produced per c.c. from n atoms (or molecules) per c.c. is nq, where

$$q = \int_{\nu_0}^{\infty} kS_\nu e^{-\tau \sec z}(h\nu)^{-1}\,d\nu,$$

where S_ν is the intensity of sunlight entering the atmosphere, τ is the optical depth in the atmosphere or $\tau = \int_x^\infty kn\,dx$ and z is the sun's zenith distance: x is the height measured upwards.

Now suppose that the atoms or molecules are distributed exponentially with height according to the law $n = n_0 e^{-x/H}$, where n_0 and H are constant, then $\tau = Hkn_0 e^{-x/H}$. Write

$$\int_{\nu_0}^{\infty} S_\nu (h\nu)^{-1}\,d\nu = S.$$

Then
$$nq = Skn_0 e^{-x/H} \exp\{-Hkn_0 \sec z\, e^{-x/H}\}.$$

Let $nq = Q$. For a given zenith distance the maximum value of Q is $Q_m(z)$ occurring at $x = x_m(z)$, where

$$x_m(z) = H \ln (Hkn_0 \sec z), \tag{1·11}$$

and if $N(z)$ is the value of n where the maximum occurs, then

$$HkN(z) = \cos z. \tag{1·12}$$

Again
$$Q_m(z) = Skn_0 \exp\{-x_m(z)/H - 1\} = S \cos z/(eH). \tag{1·13}$$

Writing $\xi = (x_m(z) - x)/H$, then

$$Q/Q_m(z) = \exp\{1 + \xi - e^{\xi}\}. \tag{1·14}$$

The quantities concerned may also be referred to $Q_m(0)$, the maximum value of Q when the sun is in the zenith, or $z = 0$. Then

$$x_m(0) = H \ln Hkn_0, \tag{1·21}$$

$$HkN_0 = 1, \tag{1·22}$$

$$Q_m(0) = S/(eH), \tag{1·23}$$

and writing $\zeta = (x_m(0) - x)/H$,

$$Q/Q_m(0) = \exp\{1 + \zeta - \sec z\, e^{\zeta}\}. \tag{1·24}$$

These expressions are due to Chapman (1931). It may be shown that they are still approximately true if $k(\nu)$ varies appreciably with ν. For example, if $k(\nu) = 0$ for $\nu < \nu_0$ and $k(\nu) = \alpha\nu^{-3}$ for $\nu > \nu_0$, these expressions are still very nearly correct in the practical case where S_ν is a black-body curve for a temperature T' and $h\nu_0/(RT') > 20$.

2. The electron density n_e can be found from Q if the absorption spectrum is the *only* method whereby electrons are produced and if the recombination coefficient α is known; for then there are αn_e^2 recombinations per c.c. per sec. and

$$\frac{dn_e}{dt} = Q - \alpha n_e^2.$$

Taking an average over 24 hr. suppose dn_e/dt zero, then, simply, $Q = \alpha n_e^2$. This expression can be reduced to a special form if the recombination is radiative, that is to say if it is the reverse of the ionization process, for there is then a relation between the absorption coefficient k and the recombination coefficient α, as was shown by Milne.

In full equilibrium at a temperature T between n_e electrons n atoms and n^+ ions per c.c. Saha's relation gives

$$\left(\frac{n_e n^+}{n}\right)_T = K(T) = 2(2\pi mRT)^{\frac{3}{2}} h^{-3} \varpi \exp(-I/RT),$$

where I is the ionization potential and ϖ is the ratio of the statistical weights of the ion and atom; m is the mass of the electron and R, h Boltzmann's and Planck's constants respectively.

Now in full equilibrium $nq_T = \alpha' n_e n^+$ $(= \alpha n_e^2)$ and in secular equilibrium $(dn_e/dt = 0)$, $nq = \alpha' n_e n^+$. Hence in secular equilibrium

$$\frac{n_e n^+}{n} = \frac{q}{q_T} K(T).$$

The value of $qK(T)/q_T$ has been investigated by Woltjer (1925) in a special case, namely, where the radiation intensity is a fraction (less than unity) of full radiation at some temperature T', while the temperature of the atmosphere (judged by the Maxwellian velocity distribution of the molecules) is T, less than T'. Woltjer's equations have been applied to stellar atmospheres by Pannekoek (1926) and to the earth's atmosphere by Bhar (1938). The subject is discussed fully by Pannekoek (1930). Let the intensity of the radiation be $(\omega/8\pi)\, B(\nu, T')$, where

$$B(\nu, T') = \frac{8\pi h\nu^3}{c^2} \frac{1}{e^{h\nu/RT'} - 1}$$

is the intensity of full radiation at the temperature T'. Then

$$\frac{n_e n^+}{n} = \omega K(T, T') = \frac{\omega}{8\pi} K(T) \times \frac{\int k(\nu)\, B(\nu, T')\, (h\nu)^{-1}\, d\nu}{\int k(\nu)\, B(\nu, T)\, (h\nu)^{-1}\, d\nu}. \tag{2·1}$$

In particular, if $k(\nu) = 0$ for $h\nu < I$ and $k(\nu) :: \nu^{-2}$ for $h\nu \gtrsim I$,

$$\omega K(T, T') = 2(2\pi m R)^{\frac{3}{2}}\, T^{\frac{1}{2}} T' \frac{\omega}{8\pi} \frac{\varpi}{h^3} \exp\left(-\frac{I}{RT'}\right). \tag{2·2}$$

If $k(\nu)$ is zero for $h\nu < I$ and is constant for $h\nu > I$, the right-hand side of (2·2) must be multiplied by

$$\left\{1 + 2\frac{RT'}{I} + 2\left(\frac{RT'}{I}\right)^2\right\} \Big/ \left\{1 + 2\frac{RT}{I} + 2\left(\frac{RT}{I}\right)^2\right\}. \tag{2·21}$$

If there are n^- negative ions per c.c., then, in place of (2·1),

$$\frac{n_e n}{n^-} = \omega K(T, T'), \tag{2·3}$$

with, of course, values of I and ϖ which differ from those in the positive-ion reaction. In the case of O^-, ϖ is $\frac{2}{3}$, and according to Vier & Mayer (1944) $I = 3{\cdot}07$ V. Then, with $T' = 6825°$, $T = 1000°$ and $\omega = 5{\cdot}41 \times 10^{-6}$,

$$n_e n / n^- = 3 \times 10^{12}.$$

Thus with $n = 3 \times 10^{10}$ and $n_e = 10^6$, representing conditions in the F_2 region, it follows that $n^- = 10^4$, giving one negative ion to about one hundred electrons.

3. Before attempting to identify the various layers of the ionosphere with particular ionization processes, one must note a difficulty which is illustrated by the following theorem. Let there be a number of ionization processes and let

ν_n be the lowest frequency capable of ionizing the atom or molecule in the nth process, numbered downwards in the order in which the maxima occurs in the atmosphere (e.g. $F_2 = 1$, $F_1 = 2$ and $E = 3$, if these regions arise from the only processes). Let S_n be the amount of light available for the nth process. Let α_n be the effective recombination coefficient in the region where the ionization due to the nth process produces a maximum electron density. *Then in an isothermal atmosphere if the quantities S_n/α_n form an increasing series, only one layer, the lowest, is observable.*

Now $$\alpha_n(N_e^2)_n = (Q_m(z))_n = S_n \cos z/(eH),$$

so that if S_n/α_n increases with decreasing height $(N_e^2)_n$ increases with decreasing height and the lowest layer has the largest electron density and reflects all the radio waves capable of reflexion by any higher layer. Hence the higher layers cannot be observed.

In the actual atmosphere the temperature probably increases with height so that H is greater for F_2 than for E. On the other hand, α decreases from about 10^{-8} at E to about 10^{-10} or 10^{-11} at F_2. With the former value (10^{-10}), αN_e^2 is roughly constant from layer to layer, and the quantities S_n must increase with increasing height in the series of observed layers.

If the solar-energy curve is at all like a black-body curve at 6000°, S_ν decreases very rapidly with increasing ν in the far ultra-violet, and a series of absorption head frequencies will give a series of values of S_n which increase rapidly with decreasing ν_n unless the absorption heads are very close together. If each layer completely cuts off the radiation capable of ionizing the next

$$S_n = \int_{\nu_n}^{\nu_{n+1}} S(\nu)\,(h\nu)^{-1}\,d\nu.$$

Values of $S_n' = 10^{-8}\beta^{-3}\int_{\nu_n}^{\nu_{n+1}} \nu^2 e^{-\beta\nu}\,d\nu$ are shown for relevant values of $\beta\nu_n$ in table 1.

TABLE 1. S_n'

$\beta\nu_n$	24	25	26	27	28
S_n'	142·3	56·7	22·5	8·93	3·53

Since for $T' = 6000°$ and $h\nu = 13{\cdot}54\,\mathrm{V}$ (the ionization potential of O), $\beta\nu = 26{\cdot}2$, successive entries in table 1 correspond to potential differences of about half a volt. A series of values of S_n' increasing with ν_n could only be obtained if successive ionization potentials closer than a quarter of a volt apart were available. In particular, the identification of two of the three layers with the reactions

$$\mathrm{O} + h\nu\,(>13{\cdot}5\,\mathrm{V}) \rightarrow \mathrm{O}^+ + e, \quad \mathrm{O}_2 + h\nu\,(>12{\cdot}5\,\mathrm{V}) \rightarrow \mathrm{O}_2^+ + e,$$

is not possible if αN_e^2 is constant—unless, of course, the hypothesis is abandoned that $S(\nu)$ does not decrease with increasing ν so rapidly as a black body at 6000° or so. If, however, $\alpha = 10^{-11}$ for the F_2 region, in which case αN_e^2, and therefore Q, is about ten times larger in both of the lower regions, S_n' for the F_2 region has only one-tenth of the value for a lower region, and from table 1, $\beta\nu = 25$ and $\beta\nu = 27$ give approximately correct values of S'_n.

According to Appleton (1937) the value of α for the F_2 region is 10^{-10}. This value depends upon afternoon minus morning differences in the F_2 critical frequencies which are attributed to recombination. According to Martyn (1945) tidal motion of charged particles is a main cause of the asymmetries in the diurnal ionization variations of the F_2 region: admitting Martyn's hypothesis, the value of α cannot be determined with any satisfaction from diurnal variations in critical frequency.

Now if $\frac{dn_e}{dt} = -\alpha n_e^2$, then $\frac{d}{dt}(\ln n_e) = -\alpha n_e$. With $n_e = 10^6$ it is seen that $-\frac{d}{dt}(\ln n_e) = 10^{-4}\,\text{sec.}^{-1}$, or (3 hr.)$^{-1}$, if $\alpha = 10^{-10}$, and (28 hr.)$^{-1}$ with $\alpha = 10^{-11}$. According to Allen's analysis (1945) of the correlations between a variety of solar and terrestrial phenomena, the F_2 region exhibits a mean period of recombination of order 1 or 2 days. This observation supports $\alpha = 10^{-11}$. The well-known lag (of the order of a day) of F_2 critical frequencies behind magnetic storms gives further support to the lower value of α. In the remainder of this paper we shall adopt $\alpha = 10^{-11}$ for the F_2 region.

4. In this section the conditions under which two observable regions could be formed by the ionization of O and of N_2 are examined, if necessary waiving the supposition that the solar-energy curve is that of a black body. Suppose, then, that the atmosphere is isothermal and let n_1, n_2 be the numbers of oxygen atoms and nitrogen molecules per c.c. Let k_1, k_2 be the continuous absorption coefficient of atomic oxygen and molecular nitrogen. Following Massey & Bates (1943), take k_1 to be independent of wave-lengths from 912 A downwards, and suppose k_2 to be independent of wave-lengths from 661 A (the beginning of the strong absorption of N_2) to shorter wave-lengths. Let

$$S_1 = \int_{\nu_1}^{\nu_2} S_\nu (h\nu)^{-1}\,d\nu \quad \text{and} \quad S_2 = \int_{\nu_2}^{\infty} S_\nu (h\nu)^{-1}\,d\nu$$

where $c\nu_1^{-1} = 912\,\text{A}$ and $c\nu_2^{-1} = 661\,\text{A}$.

Lastly let $n_1 = n_1^0 \exp(-x/H_1)$ and $n_2 = n_2^0 \exp(-x/H_2)$.

Then $H_2 = \frac{16}{28}H_1$. Hence

$$Q = S_1 k_1 n_1^0 e^{-x/H_1} \exp\{-Hk_1 n_1^0 \sec z\, e^{-x/H_1}\}$$
$$+ S_2\{k_1 n_1^0 e^{-x/H_1} + k_2 n_2^0 e^{-x/H_2}\} \exp\{-H_1 k_1 \sec z\, n_1^0 e^{-x/H_1} - H_2 k_2 \sec z\, n_2^0 e^{-x/H_2}\}.$$

Write $S_2/S_1 = S$, $k_2 n_2^0 / k_1 n_1^0 = \beta$ and $H_1 k_1 n_1^0 \sec z = A$.

Let $x/H_1 = \zeta$. Then

$$Q/(S_1 k_1 n_1^0) = \exp\{-\zeta - Ae^{-\zeta}\}[1 + S(1+\beta e^{-\frac{4}{7}\zeta}) \exp(-\tfrac{4}{7}A\beta e^{-\frac{4}{7}\zeta})]. \qquad (4\cdot1)$$

According to (4·1) two maxima of Q can occur. One may choose to reckon n_1^0 and n_2^0 at the point where one of these maxima occurs: then one maximum is at $\zeta = 0$, giving

$$0 = (A-1)\{1 + S(1+\beta)^2 \exp(-\tfrac{4}{7}A\beta)\} + S\beta(\beta + \tfrac{1}{4}) \exp(-\tfrac{4}{7}A\beta). \qquad (4\cdot21)$$

Let the other maximum occur at $\zeta = -h$ and write $Ae^h = A'$, $\beta e^{\frac{3}{4}h} = \beta'$. Then

$$0 = (A'-1)\{1+S(1+\beta')^2 \exp(-\tfrac{4}{7}A'\beta')\} + S\beta'(\beta'+\tfrac{1}{4})\exp(-\tfrac{4}{7}A'\beta'). \quad (4\cdot22)$$

The ratio of the two maxima is given by

$$Q_0/Q_{-h} = \exp\{A'-A-h\}\{1+S(1+\beta)\,e^{-\frac{4}{7}A\beta}\}\{1+S(1+\beta')\,e^{-\frac{4}{7}A'\beta'}\}^{-1}. \quad (4\cdot23)$$

From (4·21), (4·22) and (4·23) may be computed the location of the two maxima in the atmosphere, and the ratio of their intensities, given the fundamental data A, S and β. Conversely one can examine any two layers and find what values of A, S and β are required to account for them. Since nothing is known about S the latter course will be pursued.

Now from (4·21) and (4·22) both A and A' (which must be positive) are less than unity, so that for solutions which are well separate (e^h appreciably greater than 1), A must be small. Suppose as a first approximation that $A' = 1$. Then $A = e^{-h}$, and (4·21), (4·23) give a pair of equations for S and β. When the solution has been made it can be verified from (4·22) that $1-A'$ is small.

First suppose that $Q_0 = \frac{1}{10}Q_{-h}$, which may be supposed to correspond to the F_2 and F_1 regions with $\alpha = 10^{-11}$ in the upper region. No solution (with positive values) is obtained unless

$$\frac{Q_0}{Q_{-h}}\exp\{A+h-A'\} > 1,$$

so that if $Q_0/Q_{-h} = 10^{-1}$, $h > 3\cdot265$. Hence

for $h = 3\cdot401$, $\beta = 258$ and $S = 0\cdot074$,

for $h = 3\cdot912$, $\beta = 142$ and $S = 0\cdot031$.

On the other hand, suppose that the recombination coefficient is 10^{-10} in the F_2 region so that $Q_0/Q_{-h} = 1$, then

for $h = 3$, $\beta = 36\cdot7$ and $S = 0\cdot485$,

for $h = 2\cdot303$, $\beta = 19\cdot0$ and $S = 0\cdot460$.

These figures show that β is in all cases so large that unless k_2 is very much larger than k_1 (which is not likely, as the oscillator strength of the O absorption is high) the nitrogen absorption contributes mainly to the *upper* maximum.

Bhar (1938) supposes that N_2 forms the *lower* of the two regions. He is in this way able to evade the difficulty described in § 3, although he supposes that the solar-energy curve is that of a black body: but his result depends on the assumption that radiation at 661 A is not absorbed by O.

For a black body at 6000°, S is of order 10^{-4}, if the absorption heads are at 912 and 661 A.

Of two layer solutions of this type, even admitting the high values of S required, it must be remarked that they demand a rather higher ratio of nitrogen to oxygen population than seems at all likely. (Parallel calculations assuming the nitrogen to be in the atomic rather than the molecular state also give high values of β.) It may also be remarked that the solutions with $Q_0/Q_{-h} = 10^{-1}$ are open to the objection that h must be greater than 3, which corresponds to a separation of about 150 km. between the layers. Under many circumstances F_2 and F_1 come much closer together than this.

5. The lines of N_2^+ are actually observed in aurorae and in the light of the night sky, and to account for their excitation Saha (1935) postulated 'with a certain amount of hesitation' that the solar spectrum exhibits the Lyman lines of H and the principal lines of He and other elements as intensely bright lines on a much feebler continuous background. Later, Saha (1937) repeated this postulate much more emphatically, concluding that solar ultra-violet light of wave-length < 660 A is about a million times as intense as that given out by a black body at 6500°, and stating that it was difficult to see any alternative to the result.

To obtain *three* observable layers one must have the spectrum from 661 A onwards including the He line at 584 A about one-tenth as bright as the spectrum from 912 to 661 A, and that spectrum in turn half as bright as the spectrum capable of ionizing O_2 (at 12·5 V). The Lyman α and β lines will not ionize O_2 in its ground state.

Very bright lines are impossible if the solar atmosphere is in radiative equilibrium. Without going into great detail one can say roughly that if the bright lines are formed by the cyclic process described by Rosseland (1926) no bright line can emit more energy than a black-body spectrum appropriate to the colour temperature of the star integrated from the frequency of the line to infinity. (Admitting Rosseland's principle, any small volume of the atmosphere can add no more energy to a principal line than it absorbs in higher frequencies.) It is certainly not necessary to suppose that principal lines of a spectrum are bright at all (let alone intensely bright) merely because subordinate lines (in eruptions, such as Hα) appear bright.

On the other hand, the phenomenon of very high excitation in coronal lines has not yet been justified on any theory of radiative equilibrium, and in fact Saha (1945) has given a theory of the excitation of these lines by nuclear processes. It may therefore be felt that any objection to the postulate of bright lines based on the supposition that the solar envelope is in radiative equilibrium is not a fatal objection. However, in succeeding sections various mechanisms will be examined which may account for the observed ionosphere without invoking this postulate.

6. Suppose that the ionization of atomic oxygen in its ground state

$$O(^3P)+h\nu \rightarrow O^+ + e \tag{6·1}$$

forms a simple Chapman layer. Then from (1·12) $HkN(z) = \cos z$. It is seen from Massey & Bates (1943) that $k = 4{\cdot}5 \times 10^{-18}$. If it is supposed that the temperature of the atmosphere is 1000° then $H = 52{\cdot}6$ km. for oxygen, and taking $\cos z = \frac{1}{2}$ to represent average conditions, then for the number of oxygen atoms per c.c. at the height where the maximum electron production occurs

$$N = 2 \times 10^{10},$$

which is of the same order of magnitude as the number of particles per c.c. in the F region as determined by observation (Martyn & Pulley 1936). Suppose that the reaction (6·1) actually forms either the F_1 or the F_2 region, one cannot use the Pannekoek dilute equilibrium formula to get the electron density, because the radiative recombination coefficient corresponding to (6·1) is of order 10^{-12} (Massey & Bates 1943), while the observed value of α is of order 10^{-8} for the F_1 region and 10^{-10} or 10^{-11} for the F_2 region (§ 3). Since there are few negative ions per electron at such low densities (§ 2) then $n_e = n^+$ and $\alpha' = \alpha$ so far as reactions of type $O^- \rightarrow O + e$ are concerned. Massey & Bates (1943) attribute the observed high recombination to the reaction

$$O^+ + O^- \rightarrow O' + O'',$$

in which a positive ion combines with a negative ion to form two excited oxygen atoms. This reaction has so high a probability that it is important despite the comparatively low negative-ion population.

Let the number of transitions occurring per c.c. be defined as follows:

$O^+ + O^- \rightarrow O' + O''$,	$\alpha'' n^+ n^-$,	$O' + O'' \rightarrow O^+ + O^-$,	$\lambda n' n''$,
$O^- + h\nu \rightarrow O + e$,	$n^- q^-$,	$O + e \rightarrow O^- + h\nu$,	$\alpha^- n n_e$,
$O + h\nu \rightarrow O^+ + e$,	nq,	$O^+ + e \rightarrow O + h\nu$,	$\alpha' n^+ n_e$.

Then

$$\frac{dn^-}{dt} = \alpha^- n n_e - \alpha'' n^+ n^- - n^- q^- + \lambda n' n'',$$

$$\frac{dn_e}{dt} = -\alpha^- n n_e + nq + n^- q^- - \alpha' n^+ n_e.$$

Adding,

$$\frac{d}{dt}(n^- + n_e) = nq + \lambda n' n'' - \alpha' n^+ n_e - \alpha'' n^+ n^-,$$

and in secular equilibrium $n^- + n_e$ is constant, so that

$$nq + \lambda n' n'' = n^+(\alpha' n_e + \alpha'' n^-).$$

If the atmosphere is electrically neutral, $n^+ = n_e + n^-$, and if $n_e \gg n^-$ then $n^+ \doteqdot n_e$ nearly. Thus

$$nq + \lambda n' n'' = \alpha' n_e^2 + \alpha'' n_e n^-. \tag{6·2}$$

In full equilibrium the two processes balance independently, that is, $nq = \alpha' n_e^2$ and $\lambda n' n'' = \alpha'' n_e n^-$. In dilute temperature equilibrium, q will be multiplied by a dilution factor $\omega/8\pi$. If both populations of the excited states n' and n'' are of order ω in comparison with their full equilibrium values, the product $n'n''$ is of order ω^2, and $\lambda n' n''$ can be neglected in comparison with nq, giving

$$Q = \alpha n_e^2, \tag{6·3}$$

where α is now the effective recombination coefficient, equal to $\alpha' + \frac{n^-}{n_e}\alpha''$: its value may be taken to be 10^{-11} from observation.

If, however, one of the excited states is metastable, one population n' or n'' is not of order ω, and the product $n'n''$ is of the same order as q (the first order in ω), and since $\alpha'' n_e n^-$ is much greater than $\alpha' n_e^2$, $\lambda n' n''$ must be correspondingly greater than nq. The electron density will then be given by

$$\lambda n' n'' = \alpha n_e^2. \tag{6·4}$$

To investigate this point further one must examine the populations of metastable states in dilute temperature equilibrium.

7. Consider the populations of a number of states of an atom in secular equilibrium. Let n_i be the number of atoms per c.c. in the ith state and let $n_i A_{ij}$ be the number of transitions per c.c. per sec. from state i to state j. In secular equilibrium among m states

$$\sum_i n_i A_{ij} = \sum_j n_j A_{ji}, \tag{7·11}$$

both sums being from 1 to m with $i = j$ excluded. Defining A_{ii} by

$$A_{ii} + \sum_j A_{ij} = 0$$

with $j = i$ excluded from the sum, equation (7·11) becomes

$$\sum_{i=1}^{m} n_i A_{ij} = 0. \tag{7·12}$$

Let there be $(n_i)_T$ atoms per c.c. in state i in thermodynamical equilibrium at some temperature T (for a given number of atoms per c.c. in all states) and let $n_i = (n_i)_T \times t_i$. Write $(n_i)_T A_{ij} = (ij)$. Then $n_i A_{ij} = t_i(ij)$ and equation (7·12) becomes

$$\sum_{i=1}^{m} t_i(ij) = 0. \tag{7·13}$$

There are m such equations and $m-1$ unknowns (the ratios t_2/t_1, etc.) so that the determinant

$$\begin{vmatrix} (11) & (21) & \dots & (m1) \\ (12) & (22) & \dots & (m2) \\ \dots & \dots & \dots & \dots \\ (1m) & (2m) & \dots & (mm) \end{vmatrix}$$

must vanish and the quantities t_1, t_2, etc., are proportional to the minors of any row in the determinant.

In thermodynamical equilibrium $(ij) = (ji)$ and $t_1 = t_2 = \dots = t_m = 1$. Under any other circumstances, if $E_j > E_i$, (ji) is an atomic constant (for bound states) and (ij) is proportional to the intensity of the radiation promoting the transition. In dilute radiation

$$(ij) = \frac{\omega}{8\pi}(ji),$$

and in a dilute radiation field which is weakened by absorption

$$(ij) = r_{ij}\frac{\omega}{8\pi}(ji).$$

To investigate the population of a metastable state one must consider at least three atomic states: the ground state $i = 1$, the metastable state $i = 2$ and a third state $i = 3$. Then

$$\frac{t_2}{t_1} = \frac{(12).(31)+(12).(32)+(13).(32)}{(21).(31)+(21).(32)+(31).(23)} \tag{7·2}$$

$$= \frac{A\omega' r_{12} + B\omega' r_{13}}{A + B\omega' r_{23}}, \tag{7·21}$$

where $\omega' = \omega/8\pi$, $A = (21).(31)+(21).(32)$ and $B = (31).(32)$. If the transition (21) is a normal one, t_2/t_1 is of order ω', but if state 2 is metastable and A is small compared with $B\omega'$

$$\frac{t_2}{t_1} = \frac{r_{13}}{r_{23}}. \tag{7·22}$$

Thus the population of a metastable state is unity, not ω', at the top of the atmosphere (compared with its population at the solar temperature): but the metastable atoms are not distributed exponentially with height. That the population of a metastable state is of order unity in dilute equilibrium was first pointed out by Rosseland (1926).

8. Bates and Massey (1943) list eleven possible pairs of excited states O′ and O″. Of these six may show a large coefficient α''. They are

$$^1\mathrm{S}+{}^1\mathrm{S},\quad {}^3\mathrm{P}+{}^5\mathrm{S}^0,\quad {}^3\mathrm{P}+{}^3\mathrm{S}^0;\qquad {}^3\mathrm{P}+{}^5\mathrm{P},\quad {}^3\mathrm{P}+{}^3\mathrm{P},\quad {}^1\mathrm{D}+{}^5\mathrm{S}^0.$$

Of these the last three are ruled out if the value of the detachment potential is 3·0 V, as given by Vier & Mayer (1944). The states $^1\mathrm{S}$ and $^1\mathrm{D}$ are metastable. It is uncertain which pair of states gives the highest value of α'' (and λ), but since in one pair both the states are metastable, the product $\lambda n'n''$ must be greatest for that pair, namely,

$$\mathrm{O}^+ + \mathrm{O}^- \rightarrow \mathrm{O}(^1\mathrm{S}) + \mathrm{O}(^1\mathrm{S}).$$

Recognizing that more than one pair of excited states is possible one may write (6·4)

$$\Sigma\lambda_{rs} n_r n_s = \alpha n_e^2. \tag{8·1}$$

In full equilibrium $\lambda_{rs}(n_r n_s)_T = \alpha''_{rs}(n_e^2)_T$

for each pair, and writing $\lambda_{rs}(n_r n_s)_T = \gamma_{rs}\Sigma\lambda_{rs}(n_r n_s)_T$,

(8·1) becomes $$n_e^2/(n_e^2)_T = \Sigma\gamma_{rs} t_r t_s. \tag{8·2}$$

The quantities γ_{rs} are not, strictly speaking, independent of T, as the cross-sections for mutual neutralization of O^+ and O^- (the quantities α''_{rs}) are not independent of T. In writing an equilibrium value of n_e it is supposed that n is given.

Despite its approximate character, equation (8·2) exhibits the important points of the argument. If $t_r = t_s = 1$ and if the r, s pair exhibits the largest coefficient, so that γ_{rs} is nearly equal to unity, n_e will be equal to the number of electrons per c.c. in equilibrium with n oxygen atoms per c.c. *at the solar temperature* without any dilution factor. While γ_{rs} may have a high value, even approaching unity, it is doubtful whether t_r and t_s attain high values on account of absorption (the term r_{13} in (7·22)), and because in the important case (the ^1S state of O) the condition $A \ll B\omega'$ is not satisfied. If t_r and t_s are very small, equation (6·2) reduces to (6·3) rather than (6·4). The result may be expressed in terms of the solar temperature required to produce the observed electron concentration. Making no allowance for day and night averaging, and taking $n_e = 10^6$ and $n = 2 \times 10^{10}$, then (by day)

$\Sigma\gamma_{rs}t_rt_s$	$1{\cdot}7 \times 10^{-3}$	4×10^{-4}	10^{-4}	0
solar temperature	5000°	5500°	6000°	6500°

The entry 6500° for $\Sigma\gamma_{rs}t_rt_s$ zero assumes $\alpha = 10^{-11}$: other entries are unaffected by the choice of α. If $\alpha = 10^{-10}$ and the metastable states make no appreciable contribution, a temperature of 7100° is required.

If the solar atmosphere is in radiative equilibrium one would expect a lowering of the visible solar curve in the far ultra-violet, especially beyond 912 A, where one would expect very strong Lyman absorption. In the visible region radiation is received from depths considerably below, and therefore considerably hotter than, the extreme surface, which is at 4830°. An effective temperature in the Lyman continuum not much greater than 5000° might be expected. The populations of the metastable states of O are supported by radiation on the long-wave side of 912 A, so that the process $O(^1S)+O(^1S) \rightarrow O^+ + O^-$ is favoured relative to $O(^3P)+h\nu \rightarrow O^+ + e$ if the solar-energy curve exhibits an absorption continuum at 912 A.

In any event it appears possible to account for the observed electron density in the F_2 region by supposing that the electrons are supplied by the ionization of O to O^+, either directly or by the double process $O(^1S)+O(^1S) \rightarrow O^+ + O^-$ followed by $O^- \rightarrow O + e$. With $\alpha = 10^{-11}$ the solar-energy curve required does not exceed the observed black-body curve at about 6800° in the visible spectrum, and a solar curve weakened below this will produce the observed ionization if the metastable states are abundant.

9. If t remained constant (not necessarily unity) throughout the atmosphere, a metastable state would yield a higher value of Q than the ground state of the same atom, because S is greater for the metastable state. In general the absorption coefficient per gram of the atmosphere is much lower for absorption by a metastable state than for absorption by the ground state. The layer formed by the metastable state would therefore be below the layer formed by the ground state and completely screen it from observation. With the ground state distributed exponentially with height and with t constant, in fact only one layer would be seen, formed by that metastable state which gave the highest value of S. This does not, however, occur in practice, because t is not constant, as it contains the absorption term r_{13}. If r_{13} is sufficiently small, the metastable state will not form an observable layer because the population of the metastable state has not reached sufficiently large values to form a maximum electron density above the region where collisions become important. Below 100 km. or so, collisions become more important than radiative

processes in determining the populations of the metastable states. They will therefore be appropriate to the temperature of the earth's atmosphere and will not yield the full value of Q given by S/eH.

The two reactions $$\mathrm{O}(^1\mathrm{D}_2)+h\nu(11\cdot 6\,\mathrm{V})\rightarrow \mathrm{O}^+ + e,$$
$$\mathrm{N}(^2\mathrm{D}^0)+h\nu(12\cdot 1\,\mathrm{V})\rightarrow \mathrm{N}^+ + e,$$

fall into this class. When the centre of the line $1\rightarrow 3$ is blacked out, r_{13} is proportional to $n_1^{-\frac{1}{2}}$, so that n_2 is proportional to $n_1^{\frac{1}{2}}$ and still distributed exponentially. But the value of n_2 is still too small to produce an observable layer when $n_1 = 10^{13}$, and higher values of n_1 probably do not exist, as collisions inhibit the dissociation of molecules into atoms.

Metastable molecular nitrogen offers possibilities. In a homonuclear molecule, vibration bands are forbidden by the $g \rightleftharpoons u$ rule, so that all the vibrational states of the lowest electron state $x\,^1\Sigma_g^+$ cannot perform a spontaneous transition to the ground state $x\,^1\Sigma_g^+$, $v'' = 0$. For states with $v'' > 6$ the energy necessary to ionize the molecule is less than $13\cdot 5$ V, so that ionization can be accomplished by radiation which passes the barrier
$$\mathrm{O}(^3\mathrm{P})+h\nu(13\cdot 5\,\mathrm{V})\rightarrow \mathrm{O}^+ + e.$$

If $v'' > 16$ the energy of the $x\,^1\Sigma_g^+$ state is greater than that of a $^1\Pi_u$ state so that $x\,^1\Sigma_g^+, v'' > 16$ is not metastable. An observable layer may be formed by

$$\mathrm{N}_2(x\,^1\Sigma_g^+,\ 16 > v'' > 6)+h\nu(<13\cdot 5\,\mathrm{V})\rightarrow \mathrm{N}_2^+ + e. \tag{9·1}$$

Too little is known about the oscillator strengths concerned to apply a quantitative test to the hypothesis that (9·1) is responsible for the F_1 layer.

According to Mulliken (1932) molecular nitrogen should exhibit a $^1\Pi_g$ state, which would be metastable if it lay below $^1\Pi_u$. The $^1\Pi_g$ state has not been observed. The strongest unidentified band in the night-sky spectrum occurs at 3556 A. If this is the forbidden transition $^1\Pi_g\rightarrow {}^1\Sigma_g$, the state has an excitation potential of $3\cdot 5$ V and requires 12 V to become ionized. Such a metastable state would be excellently placed for forming the F_1 layer.

One may at this point discuss briefly another source of the electrons in the F_1 layer, namely nitric oxide (NO). The first ionization potential is $9\cdot 1$ V, but the ionization continuum has not been observed (Price 1943). Since, however, examination of the N_2 absorption spectrum with high dispersion by Worley & Jenkins (1938) has revealed at the first ionization potential an ionization continuum which was at one time thought not to exist, or to be very weak (Price & Collins 1935), it may be supposed that higher dispersion would reveal a continuum for NO. A single band in the spectrum of NO^+ has been observed by Duffieux & Grillet (1936, 1937). The band head is at 5999 A. If this band is a transition to the ground state, the second ionization potential of NO is at $11\cdot 2$ V.

Like every other constituent, NO would be distributed exponentially with height in an isothermal atmosphere in which there was no change in the intensity of photo-dissociating radiation. The population of NO in the actual atmosphere will accordingly be on an exponential curve at great heights at which both O_2 and N_2 are highly dissociated. The population of NO will fall off from this curve at just that point where absorption of the radiation which dissociates N_2 begins to be appreciable. This may be expected to occur somewhere near the height of the F_1 region.

10. Suppose that the electrons which form the F_1 region are supplied by absorption of light with energy below 13·5 V, the absorbing agent not being distributed exponentially with height. It is possible that several of the processes mentioned in § 9 contribute to the electron density. Radiation of energy 12·5 V will be diminished in intensity, but the diminution will not be so great as to exclude the straightforward reaction

$$O_2\,(x\,{}^3\Sigma_g,\ v''=0)+h\nu(12{\cdot}5\,\mathrm{V})\rightarrow O_2^+ + e$$

from forming the E region. As in N_2, vibration bands are forbidden and a number of vibrational levels are metastable. In addition the state ${}^1\Sigma_g$ is metastable according to Mulliken. There is no doubt that O_2 can supply enough electrons to form the E region: the only difficulty is to explain why it does not supply so many that the E region will screen the upper layers. But the E region is at a considerably higher pressure than the upper regions, and it may be supposed that collisions are no longer negligible at this level, and prevent the metastable states of O_2 from building up too high a population.

The writer acknowledges with pleasure the assistance he has derived from discussing ionospheric problems with Dr D. F. Martyn and Dr C. W. Allen.

REFERENCES

Allen 1946 *Terr. Mag.* **51**, 1.
Appleton 1937 *Proc. Roy. Soc.* A, **162**, 451.
Bates & Massey 1943 *Phil. Trans.* A, **239**, 269.
Bhar 1938 *Indian J. Phys.* **12**, 363.
Chapman 1931 *Proc. Roy. Soc.* A, **132**, 353.
Duffieux & Grillet 1936 *C.R. Acad. Sci., Paris*, **202**, 937.
Duffieux & Grillet 1937 *C.R. Acad. Sci., Paris*, **205**, 39.
Martyn 1945 *Nature*, **155**, 363.
Martyn & Pulley 1936 *Proc. Roy. Soc.* A, **154**, 453.
Massey & Bates 1943 *Rep. Progr. Phys.* **9**, 62.
Mulliken 1932 *Rev. Mod. Phys.* **4**, 53.
Pannekoek 1926 *Proc. Acad. Sci. Amst.* **29**, 1165.
Pannekoek 1930 *Handbuch der Astrophysik*, 3/1, 288–291.
Price 1943 *Rep. Progr. Phys.* **9**, 13.
Price & Collins 1935 *Phys. Rev.* **48**, 714.
Rosseland 1926 *Astrophys. J.* **63**, 218.
Saha 1935 *Proc. Nat. Inst. Sci. India*, **1**, 222.
Saha 1937 *Proc. Roy. Soc.* A, **160**, 57.
Saha 1945 *Proc. Phys. Soc.* **57**, 271.
Vier & Mayer 1944 *J. Chem. Phys.* **12**, 28.
Woltjer 1925 *Physica*, 5, 406.
Worley & Jenkins 1938 *Phys. Rev.* **54**, 304.

Stability of viscous flow between rotating cylinders. I

By D. Meksyn, D.Sc.

(*Communicated by G. Temple, F.R.S.—Received* 6 *December* 1945)

1. Introduction and summary

The problem of stability of viscous flow between two rotating cylinders was treated theoretically and experimentally by Taylor (1923, 1936).

After Taylor, the problem was considered and extended in several papers.

Faxen (1928) discussed the convergence of Taylor's expansions. Goldstein (1937) considered theoretically the case of flow between rotating cylinders under pressure. The question was also discussed by Jeffreys (1928) and Synge (1938).

The problem was dealt with experimentally by Lewis (1927), Cornish (1933) and Fage (1938).

Lord Rayleigh's criterion for an inviscid fluid, namely, that a fluid is stable at all speeds if $\Omega_2 R_2^2 > \Omega_1 R_1^2$ and unstable if $\Omega_2 R_2^2 < \Omega_1 R_1^2$, was confirmed by Taylor for the case of viscous fluid; the latter condition being, however, modified in the sense that the motion becomes unstable only if Ω_1 is greater than a certain value. (R_1, Ω_1 and R_2, Ω_2 are the radii and angular velocities of the inner and outer cylinders respectively.)

Taylor also found that in the case of instability, if the velocities are proportional, say, to $\cos \lambda z\, e^{\sigma t}$, and the ratio Ω_2/Ω_1 is positive, λ satisfies to a fairly high degree of approximation the relation $\lambda d = \pi$, d being the distance between the cylinders.

Taylor's experiments confirmed his theoretical results and showed a very close agreement between the theoretical and experimental values of the critical angular speed Ω_1.

The method of solution employed by Taylor was the expansion of the velocities in orthogonal Bessel functions of order zero and unity.

This method leads to the general solution of the problem for both cases, namely, when the cylinders rotate in the same, or in opposite directions; the formulae obtained, however, are rather complicated.

The method of solution employed in the present paper is an asymptotic expansion of the velocities in inverse powers of a large parameter.

The final formulae are very simple; the solution, however, is only valid in the case when the cylinders rotate in the same direction. The case when the cylinders rotate in opposite directions would require a separate investigation.

This solution confirmed Taylor's theoretical conclusions, and showed a complete agreement with the numerical results.

Additionally, it was found that for a fixed value of Ω_1 there is only a limited number of possible wave-lengths; this restriction is very drastic, and, since for large values of Ω_1/ν (R_1, R_2 and Ω_2/Ω_1 being fixed) the motion becomes turbulent, it is probable that only a single wave-length can be observed.

It is also shown that there are no solutions for complex or imaginary σ; there are, however, solutions for positive and negative values of σ, as was found by Taylor.

It should be emphasized that this solution is only valid for positive μ, i.e. when the cylinders rotate in the same direction.

2. EQUATIONS OF MOTION

Below are given the equations of motion as set out in Taylor's paper (1923, pp. 294–5); also Taylor's notation will be followed as closely as possible.

Let V be the velocity at any point of an incompressible viscous fluid in steady motion between two infinitely long rotating cylinders of radii R_1 and R_2 ($R_2 > R_1$).

If r is the distance of a point from the axis, then

$$V = Ar + \frac{B}{r}, \tag{2·1}$$

where A and B are connected with the angular velocities Ω_1 and Ω_2 by the relations

$$\Omega_1 = A + B/R_1^2, \quad \Omega_2 = A + B/R_2^2, \tag{2·2}$$

whence

$$A = \frac{\Omega_1(1-\mu R_2^2/R_1^2)}{1-R_2^2/R_1^2}, \quad B = \frac{R_1^2\Omega_1(1-\mu)}{1-R_1^2/R_2^2}, \tag{2·3}$$

where $\mu = \Omega_2/\Omega_1$.

Let u, $V+v$, w be the components of velocity in the disturbed motion, u along r, $V+v$ perpendicular to the meridian plane, and w is the component along the axis.

It is assumed that u, v and w are small compared with V, and that the disturbance is symmetrical, so that they are functions of r, z and t only; z is the co-ordinate parallel to the axis, and t is the time.

Neglecting products and squares of u, v, w the equations of the disturbed motion may be written

$$\left.\begin{aligned} \frac{1}{\rho}\frac{\partial p}{\partial r} - \frac{V^2}{r} &= -\frac{\partial u}{\partial t} + 2\left(A + \frac{B}{r^2}\right)v + \nu\left(\nabla_1^2 u + \frac{\partial^2 u}{\partial z^2} - \frac{u}{r^2}\right), \\ 0 &= -\frac{\partial v}{\partial t} - 2Au + \nu\left(\nabla_1^2 v + \frac{\partial^2 v}{\partial z^2} - \frac{v}{r^2}\right), \\ \frac{1}{\rho}\frac{\partial p}{\partial z} &= -\frac{\partial w}{\partial t} + \nu\left(\nabla_1^2 w + \frac{\partial^2 w}{\partial z^2}\right), \end{aligned}\right\} \tag{2·4}$$

where p represents pressure, ρ density, ν is kinematic viscosity, and

$$\nabla_1^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}.$$

The equation of continuity is

$$\frac{\partial u}{\partial r} + \frac{u}{r} + \frac{\partial w}{\partial z} = 0. \tag{2·5}$$

The six boundary conditions which must be satisfied are

$$u = v = w = 0 \quad \text{at} \quad r = R_1 \quad \text{and} \quad r = R_2. \tag{2·6}$$

Assume as a solution

$$u = u_0 \cos \lambda z e^{\sigma t}, \quad v = v_0 \cos \lambda z e^{\sigma t}, \quad w = w_0 \sin \lambda z e^{\sigma t}, \tag{2·7}$$

where u_0, v_0 and w_0 are functions of r only.

Eliminating p between the first and third equation (2·4), then

$$\left.\begin{aligned} &\frac{\partial u_0}{\partial r}+\frac{u_0}{r}+\lambda w_0 = 0, \\ &\nu\left(\nabla_1^2-\frac{1}{r^2}-\lambda^2-\frac{\sigma}{\nu}\right)v_0 = 2Au_0, \\ &\frac{\nu}{\lambda}\frac{\partial}{\partial r}\left\{\left(\nabla_1^2-\lambda^2-\frac{\sigma}{\nu}\right)w_0\right\} = -2\left(A+\frac{B}{r^2}\right)v_0-\nu\left(\nabla_1^2-\frac{1}{r^2}-\lambda^2-\frac{\sigma}{\nu}\right)u_0, \end{aligned}\right\} \tag{2·8}$$

the boundary conditions are

$$u_0 = v_0 = w_0 = 0 \quad \text{at} \quad r = R_1 \quad \text{and} \quad r = R_2. \tag{2·9}$$

Denoting for brevity

$$L = \nabla_1^2-\frac{1}{r^2}-\lambda_1^2, \quad M = \nabla_1^2-\lambda_1^2, \quad \lambda_1^2 = \lambda^2+\frac{\sigma}{\nu}, \tag{2·10}$$

the last two equations can be written as

$$\nu L v_0 = 2Au_0, \quad \frac{\nu}{\lambda}\frac{\partial}{\partial r}Mw_0 = -2\left(A+\frac{B}{r^2}\right)v_0-\nu L u_0. \tag{2·11}$$

From (2·8) and (2·11) then

$$u_0 = \frac{\nu}{2A}Lv_0, \quad w_0 = -\frac{1}{2A}\frac{\nu}{\lambda}\left(\frac{\partial}{\partial r}Lv_0+\frac{1}{r}Lv_0\right), \tag{2·12}$$

whence substituting in the last equation (2·11)

$$\frac{1}{2A}\frac{\nu^2}{\lambda^2}\frac{\partial}{\partial r}M\left[\frac{\partial}{\partial r}Lv_0+\frac{1}{r}Lv_0\right] = 2\left(A+\frac{B}{r^2}\right)v_0-\frac{\nu^2}{2A}L^2v_0. \tag{2·13}$$

It should be borne in mind that L and M are operators, whence the order in which they follow is essential.

Making use of (2·10) one finds the equation for v_0

$$\begin{aligned} &\frac{\partial^6 v}{\partial r^6}+\frac{3}{r}\frac{\partial^5 v}{\partial r^5}-\left(\frac{6}{r^2}+2\lambda_1^2+\lambda^2\right)\frac{\partial^4 v}{\partial r^4}+\left(\frac{12}{r^3}-\frac{4\lambda_1^2+2\lambda^2}{r}\right)\frac{\partial^3 v}{\partial r^3} \\ &\quad+\left(-\frac{27}{r^4}+\frac{6\lambda_1^2+3\lambda^2}{r^2}+\lambda_1^4+2\lambda^2\lambda_1^2\right)\frac{\partial^2 v}{\partial r^2}+\left(\frac{45}{r^5}-\frac{6\lambda_1^2+3\lambda^2}{r^3}+\frac{\lambda_1^4+2\lambda^2\lambda_1^2}{r}\right)\frac{\partial v}{\partial r} \\ &\quad+\left[-\frac{45}{r^6}+\frac{6\lambda^2+3\lambda_1^2}{r^4}-\frac{\lambda_1^4+2\lambda^2\lambda_1^2}{r^2}-\lambda^2\lambda_1^4-\frac{4\lambda^2}{\nu^2}\left(A^2+\frac{AB}{r^2}\right)\right]v = 0, \end{aligned} \tag{2·14}$$

where for the sake of brevity the subscript of v was dropped.

3. INTEGRATION OF EQUATIONS OF MOTION

Before integrating the equation (2·14) it would be convenient to express all lengths in units of R_0, the mean radius; the large parameter would then be equal to λR_0, which is an abstract number. This transformation can be avoided, however, by assuming that r remains of the order of unity, and that λ is large.

The integrals of (2·14) must now be expanded in inverse powers of λ.

Assume that

$$v = Ce^{\xi} \tag{3·1}$$

is a solution of (2·14), where ξ is a function of r containing also the large parameter, and C is a slowly varying function of r.

The order of magnitude of the above functions and their derivatives is as follows:

$$\left.\begin{matrix} \lambda, & \xi', & \xi'', & \dots, & r, & C, & C', & \dots, \\ 1, & 1, & 1, & \dots, & 0, & 0, & 0, & \dots, \end{matrix}\right\} \tag{3·2}$$

where primes denote differentiation with respect to r.

Differentiating (3·1) with respect to r, and substituting in (2·14) we find

$$\begin{aligned} &C\xi'^6 + 6C'\xi'^5 + 15C\xi'^4\xi'' + \dots + \frac{3}{r}\{C\xi'^5 + \dots\} \\ &\quad - \left(\frac{6}{r^2} + 2\lambda_1^2 + \lambda^2\right)\{C\xi'^4 + 4C'\xi'^3 + 6C\xi'^2\xi'' + \dots\} \\ &\quad + \left(\frac{12}{r^3} - \frac{4\lambda_1^2 + 2\lambda^2}{r}\right)\{C\xi'^3 + \dots\} + \left(-\frac{27}{r^4} + \frac{6\lambda_1^2 + 3\lambda^2}{r^2} + \lambda_1^4 + 2\lambda^2\lambda_1^2\right) \\ &\quad \times\{C\xi'^2 + C\xi'' + 2C'\xi' + \dots\} + \left(\frac{45}{r^5} - \frac{6\lambda_1^2 + 3\lambda^2}{r^3} + \frac{\lambda_1^4 + 2\lambda^2\lambda_1^2}{r}\right)(C\xi' + C') \\ &\quad + \left[-\frac{45}{r^6} + \frac{6\lambda^2 + 3\lambda_1^2}{r^4} - \frac{\lambda_1^4 + 2\lambda^2\lambda_1^2}{r^2} - \lambda^2\lambda_1^4 - \frac{4\lambda^2}{\nu^2}\left(A^2 + \frac{AB}{r^2}\right)\right]C = 0. \end{aligned} \tag{3·3}$$

Collecting terms of the order of magnitude of six and five respectively, and equating them to zero, then

$$\left.\begin{aligned} Cf(\xi') &= 0, \\ C'\frac{\partial f(\xi')}{\partial \xi'} + \frac{1}{2}\frac{C}{r}\frac{\partial f(\xi')}{\partial \xi'} + \frac{C}{2}\frac{\partial^2 f(\xi')}{\partial \xi'^2}\xi'' &= 0, \\ f(\xi') = \xi'^6 - (2\lambda_1^2 + \lambda^2)\xi'^4 + (\lambda_1^4 + 2\lambda^2\lambda_1^2)\,\xi'^2 - \lambda^2\lambda_1^4 &- \frac{4\lambda^2}{\nu^2}\left(A^2 + \frac{AB}{r^2}\right). \end{aligned}\right\} \tag{3·4}$$

It was assumed in the above that B/r^2 is of the same order as A, since B contains R_1^2 as a factor, and r changes between R_1 and R_2, the difference between them being small in comparison with each of them; A was assumed to be of the order of λ^2, which will be justified *a posteriori*.

From the first equation (3·4) we find ξ', and integrating the second equation, then

$$C = \frac{\text{const.}}{r^{\frac{1}{2}}[\partial f(\xi')/\partial \xi']^{\frac{1}{2}}}. \tag{3·5}$$

The above solution represents only the dominant term; the contribution of the next term is negligible.

The solution of the equation may now be completed. From (3·4) it is found that

$$f(\xi') = (\xi'^2-\lambda^2)(\xi'^2-\lambda_1^2)^2 - \frac{4\lambda^2}{\nu^2}\left(A^2+\frac{AB}{r^2}\right) = 0. \tag{3·6}$$

It is assumed in what follows, unless otherwise stated, that $\sigma = 0$, i.e. $\lambda_1 = \lambda$, whence

$$\xi'^2 = \lambda^2\left\{1+\omega\left[\frac{4}{\nu^2\lambda^4}\left(A^2+\frac{AB}{r^2}\right)\right]^{\frac{1}{3}}\right\}, \tag{3·7}$$

where

$$\omega^3 = 1, \tag{3·8}$$

ω being any cube root of one.

Since our integrals will only be needed to solve the boundary conditions equation, further simplifications can be made in (3·6).

Put $$R_2-R_1 = 2r_0, \quad R_2+R_1 = 2R_0; \quad y = r-R_0, \quad d = 2r_0. \tag{3·9}$$

It can be easily shown from (2·3) and (3·9) that

$$\left.\begin{aligned} A+\frac{B}{r^2} &= \Omega_1\left[\frac{\left(1-\frac{3}{2}\frac{r_0}{R_0}\right)+\mu\left(1+\frac{3}{2}\frac{r_0}{R}\right)}{2}\right]\left\{1+\alpha\left(-\frac{y}{2r_0}+\frac{3y^2}{4r_0R_0}\right)\right\}, \\ \alpha &= \frac{2(1-\mu)}{(1+\mu)-\frac{3}{2}\frac{r_0}{R_0}(1-\mu)}, \end{aligned}\right\} \tag{3·10}$$

where only terms up to r_0/R_0 were retained; whence

$$\left.\begin{aligned} \frac{4}{\nu^2\lambda^4}\left(A^2+\frac{AB}{r^2}\right) &= h^3\left[1+\alpha\left(-\frac{y}{2r_0}+\frac{3y^2}{4r_0R_0}\right)\right], \\ h^3 &= \frac{4\Omega_1^2}{\nu^2\lambda^4}\frac{1-\mu R_2^2/R_1^2}{1-R_2^2/R_1^2}\frac{(1+\mu)-\frac{3}{2}\frac{r_0}{R_0}(1-\mu)}{2}, \end{aligned}\right\} \tag{3·11}$$

where α is given by (3·10); it can therefore be assumed that

$$\xi'^2 = \lambda^2\left\{1+\omega h\left[1+\alpha\left(-\frac{y}{2r_0}+\frac{3y^2}{4r_0R_0}\right)\right]^{\frac{1}{3}}\right\}. \tag{3·12}$$

To find ξ one has to integrate ξ' (3·12) with respect to y between zero and y; in the characteristic equation there will appear terms like

$$\int_{R_1}^{R_2}\xi' dr = \int_{-r_0}^{r_0}\xi' dy,$$

and it is clear that the term in y^2 in ξ will make no contribution; the term in y in (3·12) can therefore be disregarded in the first approximation, and it be assumed that ξ' is constant within the range considered; it does not follow of course that, generally, ξ'' can be taken equal to zero; whence

$$\xi' = \pm\lambda(1+\omega h)^{\frac{1}{2}}, \quad \xi = \xi' r. \tag{3·13}$$

Now introduce the following notation; let

$$\left.\begin{aligned} \omega_1 = \exp\left(\frac{2\pi i}{3}\right), \quad \omega_2 = \exp\left(-\frac{2\pi i}{3}\right) \\ a = (1+\omega_1 h)^{\frac{1}{2}}, \quad \bar{a} = (1+\omega_2 h)^{\frac{1}{2}}, \quad a_0 = (1+h)^{\frac{1}{2}} \end{aligned}\right\} \tag{3·14}$$

where $\bar{a}$ can be taken as the complex conjugate of a; whence

$$\left.\begin{aligned} \xi_1' = \lambda a, \quad \xi_2' = \lambda\bar{a}, \quad \xi_3' = -\lambda a, \\ \xi_4' = -\lambda\bar{a}, \quad \xi_5' = \lambda a_0, \quad \xi_6' = -\lambda a_0 \end{aligned}\right\} \tag{3·15}$$

and the six integrals of (2·14) are, to the first approximation,

$$\left.\begin{aligned} v_1 &= \frac{1}{r^{\frac{1}{2}} f'_{\xi_1'}(\xi_1')^{\frac{1}{2}}}\exp(\lambda a r), \quad v_2 = \frac{1}{r^{\frac{1}{2}} f'_{\xi'}(\xi_2')^{\frac{1}{2}}}\exp(\lambda\bar{a}r), \\ v_3 &= \frac{1}{r^{\frac{1}{2}} f'_{\xi_3'}(\xi_3')^{\frac{1}{2}}}\exp(-\lambda a r), \quad v_4 = \frac{1}{r^{\frac{1}{2}} f'_{\xi_4'}(\xi_4')^{\frac{1}{2}}}\exp(-\lambda\bar{a}r), \\ v_5 &= \frac{1}{r^{\frac{1}{2}} f'_{\xi_5'}(\xi_5')^{\frac{1}{2}}}\exp(\lambda a_0 r), \quad v_6 = \frac{1}{r^{\frac{1}{2}} f'_{\xi_6'}(\xi_6')^{\frac{1}{2}}}\exp(-\lambda a_0 r). \end{aligned}\right\} \tag{3·16}$$

The differentiation of $f(\xi')$ is taken with respect to ξ' and not r; in what follows the ξ''s can be taken as constant and the $f'(\xi')$ disregarded.

4. BOUNDARY CONDITIONS

From (2·9) and (2·12) it is found that the boundary conditions are

$$v_0 = 0, \quad Lv_0 = 0, \quad \frac{\partial}{\partial r}Lv_0 = 0 \tag{4·1}$$

at $r = R_1$ and $r = R_2$, whence from (2·10) the above conditions become

$$v_0 = 0, \quad \frac{d^2v_0}{dr^2} = 0, \quad \frac{d^3v_0}{dr^3} - \lambda_1^2\frac{dv_0}{dr} = 0 \tag{4·2}$$

at $r = R_1$ and $r = R_2$, where the terms $dv_0/r\,dr$ and dv_0/r^2dr were disregarded in the second and third equations (4·2) as being of lower order of magnitude.

A further simplification can be made in the expressions (3·16) before they are used in the boundary conditions.

The factors $f'(\xi')$ in the denominators (3·16) can be disregarded, since they appear symmetrically in the characteristic equation, and drop out from the final expression; and, since the terms due to ξ'' were found to be very small in comparison with the terms due to ξ', the latter can be treated as constant in the differentiations.

Accordingly it is assumed that

$$\left.\begin{aligned} v_1 &= \frac{1}{r^{\frac{1}{2}}}\exp(a\lambda r), & v_2 &= \frac{1}{r^{\frac{1}{2}}}\exp(\bar{a}\lambda r), \\ v_3 &= \frac{1}{r^{\frac{1}{2}}}\exp(-a\lambda r), & v_4 &= \frac{1}{r^{\frac{1}{2}}}\exp(-\bar{a}\lambda r), \\ v_5 &= \frac{1}{r^{\frac{1}{2}}}\exp(a_0\lambda r), & v_6 &= \frac{1}{r^{\frac{1}{2}}}\exp(-a_0\lambda r). \end{aligned}\right\} \tag{4·3}$$

It will be necessary to differentiate the v's when substituting in the boundary conditions (4·2); the dominant term only will be retained in these differentiations; for instance,

$$\frac{\partial v_1}{\partial r} = \frac{a\lambda}{r^{\frac{1}{2}}}\exp(a\lambda r) - \frac{1}{2r^{\frac{3}{2}}}\exp(a\lambda r),$$

and the second term on the right-hand side will be disregarded since it is of the order $1/a\lambda r$, which was assumed to be small.

It should be noticed that, even if it were desirable to get a better approximation, it would still be simpler to disregard these terms in the first approximation and introduce them only in the second approximation.

Let

$$\left.\begin{gathered} \epsilon_1 = \epsilon_3 = a^2\lambda^2, \quad \epsilon_2 = \epsilon_4 = \bar{a}^2\lambda^2, \quad \epsilon_5 = \epsilon_6 = a_0^2\lambda^2, \\ \eta_1 = -\eta_3 = a\lambda^3\left(a^2 - \frac{\lambda_1^2}{\lambda^2}\right), \quad \eta_2 = -\eta_4 = \bar{a}\lambda^3\left(\bar{a}^2 - \frac{\lambda_1^2}{\lambda^2}\right), \\ \eta_5 = -\eta_6 = a_0\lambda^3\left(a_0^2 - \frac{\lambda_1^2}{\lambda^2}\right). \end{gathered}\right\} \tag{4·4}$$

In the present case $\lambda_1 = \lambda$.

Let now the general solution of (2·14) be

$$v_0 = \sum_{n=1}^{6} A_n v_n, \tag{4·5}$$

where $A_1, \ldots, A_6$ are constants.

The boundary conditions (4·2) lead, by making use of (4·4), to the following six equations:

$$\left.\begin{aligned} \sum_{n=1}^{6} A_n v_n^{(1)} &= 0, & \sum_{n=1}^{6} A_n \epsilon_n v_n^{(1)} &= 0, & \sum_{n=1}^{6} A_n \eta_n v_n^{(1)} &= 0, \\ \sum_{n=1}^{6} A_n v_n^{(2)} &= 0, & \sum_{n=1}^{6} A_n \epsilon_n v_n^{(2)} &= 0, & \sum_{n=1}^{6} A_n \eta_n v_n^{(2)} &= 0, \end{aligned}\right\} \tag{4·6}$$

where $v_n^{(1)}$ and $v_n^{(2)}$ denote the values of v_n at $r = R_1$ and $r = R_2$ respectively.

Eliminating the constants $A_1, \ldots, A_6$ between (4·6) the following characteristic equation is found:

$$\Delta = \begin{vmatrix} v_1^{(1)} & \cdots & v_6^{(1)} \\ \epsilon_1 v_1^{(1)} & \cdots & \epsilon_6 v_6^{(1)} \\ \eta_1 v_1^{(1)} & \cdots & \eta_6 v_6^{(1)} \\ \cdots & \cdots & \cdots \\ \eta_1 v_1^{(2)} & \cdots & \eta_6 v_6^{(2)} \end{vmatrix} = 0. \tag{4·7}$$

5. EVALUATION OF THE DETERMINANT

To evaluate the determinant one must bear in mind the order of magnitude of its terms.

The constant a_0 may be either real or purely imaginary, but a and $\bar{a}$ are complex, the real part of which can be assumed as positive.

In the determinant there will be terms proportional to

$$\left.\begin{aligned} &v_1^{(2)} v_2^{(2)} v_3^{(1)} v_4^{(1)} v_5^{(2)} v_6^{(1)} \sim \exp\{(a+\bar{a})\lambda d + \lambda a_0 d\}, \\ &v_1^{(2)} v_2^{(2)} v_3^{(2)} v_4^{(1)} v_5^{(1)} v_6^{(1)} \sim \exp(\lambda \bar{a} d), \\ &v_1^{(1)} v_2^{(2)} v_3^{(1)} v_4^{(2)} v_5^{(2)} v_6^{(1)} \sim \exp(\lambda a_0 d), \end{aligned}\right\} \tag{5·1}$$

and, since $\exp(a\lambda d)$ is large, the terms of the second and third kind (5·1) will be much smaller; whence they can be disregarded.

The remaining terms lead to

$$\left.\begin{aligned} &\Delta = P \sinh \lambda a_0 d \cosh \lambda(a+\bar{a}) d + Q \cosh \lambda a_0 d \sinh \lambda(a+\bar{a}) d = 0, \\ &P = [\Delta_{34} + \epsilon_6(\eta_3 - \eta_4)][\Delta_{12} + \epsilon_5(\eta_1 - \eta_2)] + \eta_5 \eta_6 (\epsilon_1 - \epsilon_2)(\epsilon_3 - \epsilon_4), \\ &Q = \eta_5(\epsilon_3 - \epsilon_4)[\Delta_{12} + \epsilon_6(\eta_1 - \eta_2)] + \eta_6(\epsilon_1 - \epsilon_2)[\Delta_{34} + \epsilon_5(\eta_3 - \eta_4)], \\ &\Delta_{\mu\nu} = \epsilon_\mu \eta_\nu - \epsilon_\nu \eta_\mu. \end{aligned}\right\} \tag{5·2}$$

To evaluate P and Q they are divided out by $(\epsilon_1 - \epsilon_2)(\epsilon_3 - \epsilon_4)$. Inserting in P and Q the expressions for Δ_{12} and Δ_{34} it is found that, after few transformations,

$$\left.\begin{aligned} \frac{P}{(\epsilon_1 - \epsilon_2)(\epsilon_3 - \epsilon_4)} &= -\left\{\left[\frac{\epsilon_2 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_1 - \frac{\epsilon_1 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_2\right]^2 + \eta_5^2\right\}, \\ \frac{Q}{(\epsilon_1 - \epsilon_2)(\epsilon_3 - \epsilon_4)} &= -2\eta_5\left[\frac{\epsilon_2 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_1 - \frac{\epsilon_1 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_2\right]^2, \end{aligned}\right\} \tag{5·3}$$

whence (5·2) becomes

$$\left\{\left[\frac{\epsilon_2 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_1 - \frac{\epsilon_1 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_2\right]^2 + \eta_5^2\right\} \tanh \lambda a_0 d + 2\eta_5 \left\{\frac{\epsilon_2 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_1 - \frac{\epsilon_1 - \epsilon_5}{\epsilon_1 - \epsilon_2}\eta_2\right\} \tanh \lambda(a+\bar{a}) d = 0. \tag{5·4}$$

Making use of (4·4) then (3·14)

$$\frac{\epsilon_2 - \epsilon_5}{\epsilon_1 - \epsilon_2} = \omega_1, \quad \frac{\epsilon_1 - \epsilon_5}{\epsilon_1 - \epsilon_2} = -\omega_2, \quad \eta_1 = a\lambda^3 h \omega_1, \quad \eta_2 = \bar{a}\lambda^3 h \omega_2, \quad \eta_5 = a_0 \lambda^3 h, \tag{5·5}$$

whence (5·4) becomes

$$\{(a\omega_1^2+\bar{a}\omega_2^2)^2+a_0^2\}\tanh\lambda d(1+h)^{\frac{1}{2}}+2a_0(a\omega_1^2+\bar{a}\omega_2^2)\tanh(a+\bar{a})\lambda d=0, \tag{5·6}$$

where the symbols employed in (5·6) are explained in (2·7) and (3·14), and $d = R_2 - R_1$ is the distance between the cylinders.

6. Solution of the characteristic equation, $\sigma = 0$

(i) $1+h>0$. Since $(1+h)^{\frac{1}{2}}\lambda d$ is large, at least equal to two, and $a+\bar{a}$ is positive, one can put in (5·6)

$$\tanh\lambda a_0 d=1,\quad \tanh\lambda(a+\bar{a})\,d=1, \tag{6·1}$$

whence, extracting the square root (5·6),

$$a\omega_2+\bar{a}\omega_1+a_0=0. \tag{6·2}$$

Transferring a_0 to the right-hand side, and squaring (3·14)

$$2a\bar{a}=2(1-h+h^2)^{\frac{1}{2}}=2(1+h), \tag{6·3}$$

whence, by squaring again, one obtains

$$h=0. \tag{6·4}$$

This value of h, however, reduces (4·3) to two asymptotically independent solutions only, whence it has to be discarded.

The following important result is then obtained; in order that the motion may be unstable it is necessary that h shall be negative, its absolute value being larger than unity; i.e. from (3·11) the condition of instability is

$$\left|\frac{4\Omega_1^2}{\nu^2\lambda^4}\,\frac{1-\mu R_2^2/R_1^2}{1-R_2^2/R_1^2}\,\frac{1+\mu-\dfrac{3}{2}\dfrac{r_0}{R_0}(1-\mu)}{2}\right|>1, \tag{6·5}$$

i.e. for fixed Ω_1, μ, R_1 and R_2, λ *has an upper limit*; *there is*, therefore, only *a limited number of possible solutions*.

This is in striking contrast to the case of elastic vibrations where there is always an infinite number of harmonics.

(ii) $1+h<0$. Putting $$(1+h)^{\frac{1}{2}}=i|1+h|^{\frac{1}{2}}, \tag{6·6}$$

let (3·14) $$a=(1+\omega_1 h)=Ke^{ik}, \tag{6·7}$$

and (5·6) becomes, after few transformations,

$$\tan|1+h|^{\frac{1}{2}}\lambda d=-\frac{2|1+h|^{\frac{1}{2}}K\cos\left(k-\dfrac{2\pi}{3}\right)}{(1-h+h^2)^{\frac{1}{2}}}. \tag{6·8}$$

Since, however, $$K=(1-h+h^2)^{\frac{1}{4}}, \tag{6·9}$$

then, finally, $$\tan|1+h|^{\frac{1}{2}}\lambda d=-\frac{2|1+h|^{\frac{1}{2}}\cos\left(k-\dfrac{2\pi}{3}\right)}{(1-h+h^2)^{\frac{1}{4}}}. \tag{6·10}$$

The equation (6·10) will evidently have a solution for real values of λ; hence follows the well-known result that the motion will be unstable if (3·11)

$$1-\mu R_2^2/R_1^2>0. \tag{6·11}$$

The two equations (6·10) and (3·11) must be solved for the three unknown quantities h, λ and Ω_1, under the additional condition that Ω_1 is minimum.

Since the right-hand side of (6·10) is of the order of unity, a change in h will not affect greatly the value of the left-hand side of (6·10); whence, roughly, it is necessary to find the minimum of Ω_1, where (3·11)

$$\Omega_1 \sim h^{\frac{3}{2}}\lambda^2, \tag{6·12}$$

under the condition that (6·10)

$$|\,1+h\,|^{\frac{1}{2}}\,\lambda = \text{const.}, \tag{6·13}$$

whence the required solution is, in the first approximation,

$$h=-3.$$

To find a better approximation one assumes for h a set of values starting, say, with $h=-3$, and from (6·10) follows the corresponding values of $\lambda d = p$ (say).

Since (3·11) Ω_1 is proportional to $h^{\frac{3}{2}}\lambda^2$, i.e. to $h^{\frac{3}{2}}p^2$, the latter expression is evaluated.

Plotting $h^{\frac{3}{2}}p^2$ against h, h is found corresponding to the minimum of $h^{\frac{3}{2}}p^2$, i.e. of Ω_1, which is the required solution for h; the values of λ and Ω_1 minimum are then obtained from (6·10) and (3·11) respectively.

These computations lead to results which are in very close agreement with Taylor's, namely,

$$h=-2{\cdot}60, \quad \lambda d=\pi, \tag{6·14}$$

whence
$$\frac{4\Omega_1^2}{\nu^2\lambda^4}\,\frac{1-\mu R_2^2/R_1^2}{1-R_2^2/R_1^2}\,\frac{1+\mu-\dfrac{3}{2}\dfrac{r_0}{R_0}(1-\mu)}{2}=h^3=-17{\cdot}576. \tag{6·15}$$

If in our expression (6·15) λ is eliminated by making use of (6·14) and the term depending on $3r_0/2R_0$ is disregarded, the expression becomes identical but inverse to Taylor's expression, $-P(1-\mu)/(1+\mu)$ (Taylor 1923, p. 309), i.e.

$$-P\frac{1-\mu}{1+\mu}=\frac{1}{h^3}=-0{\cdot}0569,$$

as against Taylor's value of $-0{\cdot}0571$.

The equation (6·15) gives critical values of Ω_1 which are almost in complete agreement with Taylor's numerical results for positive μ including $\mu=0$, even for the larger values of r_0/R_0.

It should be noticed that the curve of critical Ω_1 against λ is extremely flat at the bottom, and that an increase of ½ % in Ω_1 is sufficient to change the value of the wave-length by about 5 %.

That might explain the scattering in the distances of the vortices observed by Taylor; it was, in fact, of the above order of magnitude.

7. Solution of the characteristic equation, $\sigma \neq 0$

In the above investigation it was assumed that $\sigma = 0$; consideration will now be given to the more general case when $\sigma \neq 0$.

To solve the equation

$$\xi'^6 - (2\lambda_1^2+\lambda^2)\,\xi'^4 + (\lambda_1^4 + 2\lambda^2\lambda_1^2)\,\xi'^2 - \left[\lambda^2\lambda_1^4 + \frac{4\lambda^2}{\nu^2}\left(A^2+\frac{AB}{r^2}\right)\right] = 0, \tag{7·1}$$

where

$$\lambda_1^2 = \lambda^2 + \frac{\sigma}{\nu},$$

assume

$$\xi'^2 = y + \frac{2\lambda_1^2+\lambda^2}{3}, \tag{7·2}$$

where y in (7·2) should not be confused with the one used previously; whence

$$y^3 - \frac{(\lambda_1^2-\lambda^2)}{3}y + \frac{2(\lambda_1^2-\lambda^2)^3}{27} - \frac{4\lambda^2}{\nu^2}\left(A^2+\frac{AB}{r^2}\right) = 0. \tag{7·3}$$

Put

$$y = \frac{\lambda_1^2-\lambda^2}{3}x, \tag{7·4}$$

and then, finally,

$$\xi'^2 = \lambda^2\{1+(x+2)\gamma\}, \quad \gamma = \frac{\sigma}{3\nu\lambda^2}, \tag{7·5}$$

where x is any of the roots of

$$x^3 - 3x + 2 - H = 0, \quad H = \frac{h^3}{\gamma^3}. \tag{7·6}$$

Consider the following cases.

(*a*) *σ real and small.* Consider the case when (7·5), (7·6)

$$\gamma \ll 1, \quad \text{or} \quad \frac{h^3}{\gamma^3} \gg 1. \tag{7·7}$$

There is no difficulty in discussing the more general case, but it is not needed for the present purpose.

From (7·6) one obtains, to a first approximation,

$$x = \omega\frac{h}{\gamma}, \quad \omega^3 = 1, \tag{7·8}$$

whence

$$\left.\begin{aligned} \xi'^2 &= \lambda_0^2(1+\omega h_0), \\ \lambda_0 = \lambda(1+\gamma), \quad h_0 &= \frac{h}{1+2\gamma}. \end{aligned}\right\} \tag{7·9}$$

The new characteristic equation will, therefore, be similar to (6·10) except that, instead of λ and h, one obtains λ_0 and h_0; the corresponding conditions for the critical value of Ω_1 is given by (6·14):

$$h_0 = -2{\cdot}60, \quad \lambda_0 d = \pi. \tag{7·10}$$

Denoting the new critical value of Ω_1 by Ω_0, then (3·11)

$$\frac{\Omega_0^2}{\lambda_0^4} = \frac{\Omega_1^2}{\lambda^4}, \tag{7·11}$$

and (7·9)

$$\frac{\Omega_0}{\Omega_1} = (1+\gamma)^2, \tag{7·12}$$

whence

$$\left.\begin{aligned} &\text{if } \Omega_0 > \Omega_1; \quad \gamma > 0, \sigma > 0, \\ &\text{if } \Omega_0 < \Omega_1; \quad \gamma < 0, \sigma < 0, \end{aligned}\right\} \tag{7·13}$$

i.e. the motion becomes unstable or stable according as the critical speed is increased or lowered.

(*b*) *σ complex*. Denoting as before (3·15)

$$\xi_1' = \lambda a, \quad \xi_2' = \lambda \bar{a}, \quad \ldots,$$

where the corresponding values of a, $\bar{a}$, ... are given by (7·5), it is necessary to distinguish between two cases.

(i) The constants a, $\bar{a}$ and a_0 are complex, so that

$$\tanh \lambda a_0 d = 1, \quad \tanh \lambda(a+\bar{a}) d = 1. \tag{7·14}$$

Since it leads to a negative result, only a brief sketch of this investigation will be given.

The characteristic equation (5·4) becomes

$$\frac{\epsilon_2-\epsilon_5}{\epsilon_1-\epsilon_2}\eta_1 - \frac{\epsilon_1-\epsilon_5}{\epsilon_1-\epsilon_2}\eta_2 + \eta_5 = 0. \tag{7·15}$$

Making use of the notation (3·15), then (7·5), (7·6)

$$a^2 = 1+(x_1+2)\gamma, \quad \bar{a}^2 = 1+(x_2+2)\gamma, \quad a_0^2 = 1+(x_0+2)\gamma, \tag{7·16}$$

where x_1, x_2 and x_0 are the roots of (7·6); the equation (7·15) is transformed into (4·4)

$$\frac{a}{a_0}(x_1-1)(x_2-x_0) - \frac{\bar{a}}{a_0}(x_2-1)(x_1-x_0) = -(x_0-1)(x_1-x_2). \tag{7·17}$$

To eliminate the radicals a/a_0 and $\bar{a}/a_0$ one proceeds as follows.

Making use of (4·4) and (7·16) then

$$\begin{aligned} &\frac{a^2}{a_0^2}(x_1-1)^2(x_2-x_0)^2 - \frac{\bar{a}^2}{a_0^2}(x_2-1)^2(x_1-x_0)^2 \\ &\qquad = \frac{1+(x_1+2)\gamma}{1+(x_0+2)\gamma}(x_1-1)^2(x_2-x_0)^2 - \frac{1+(x_2+2)\gamma}{1+(x_0+2)\gamma}(x_2-1)^2(x_1-x_0)^2. \end{aligned} \tag{7·18}$$

Dividing (7·18) by (7·17), and adding the obtained equation to (7·17), one eliminates $\bar{a}/a_0$; squaring the result obtained then, after few transformations,

$$\begin{aligned} &2\{(x_1-1)^2(x_2-x_0)^2 + (x_2-1)^2(x_1-x_0)^2 + \gamma H[(x_2-x_0)^2 + (x_1-x_0)^2]\} \\ &\quad \times \{(x_0-1)^2(x_1-x_2)^2 + \gamma H(x_1-x_2)^2\} = \{(x_1-1)^2(x_2-x_0)^2 - (x_2-1)^2(x_1-x_0)^2 \\ &\qquad + \gamma H[(x_2-x_0)^2 - (x_1-x_0)^2]\} + \{(x_0-1)^2(x_1-x_2)^2 + \gamma H(x_1-x_2)^2\}^2, \end{aligned} \tag{7·19}$$

where use was made of $$(x-1)^2(x+2) = H. \tag{7·20}$$

The equation (7·6) can be solved in circular functions, namely,

$$\left.\begin{aligned} x_0 = 2\sin\phi, \quad x_1 = 2\sin\left(\phi+\frac{2\pi}{3}\right) \\ x_2 = 2\sin\left(\phi-\frac{2\pi}{3}\right), \quad \text{where} \quad \sin 3\phi = \frac{2-H}{2}. \end{aligned}\right\} \tag{7·21}$$

Substituting (7·21) in (7·19), then

$$\cos\phi = 0. \tag{7·22}$$

Disregarding the factor $\cos\phi$, transferring the second term on the right-hand side in (7·19) to the left, and joining it with the first term, then, after somewhat lengthy but simple transformations,

$$(\tfrac{4}{3}+2\gamma-\tfrac{2}{3}H)+(-\tfrac{16}{3}-8\gamma+\tfrac{8}{3}\gamma H)\sin\phi+(\tfrac{16}{3}+8\gamma)\sin^2\phi = 0. \tag{7·23}$$

Since the equation (7·23) has to be satisfied by three values of ϕ, it vanishes identically, whence

$$\tfrac{4}{3}+2\gamma = 0, \quad H = 0. \tag{7·24}$$

From the solutions (7·22) and (7·24) it follows that two roots of the equation (7·6) are equal, which makes two pairs of integrals of our differential equation asymptotically equal; whence these roots have to be discarded.

There is, therefore, no solution for complex values of σ in this case.

(ii) One of the α's, say, α_0 is again purely imaginary, whence $\tanh\lambda\alpha_0 d$ is a circular function.

Since α_0 is imaginary, α_0^2 is negative, whence (7·16)

$$(x_0+2)\gamma = -l, \tag{7·25}$$

where l is positive.

Eliminating x_0 between (7·25) and (7·6), then

$$9\gamma^2+6l\gamma+l^2+\frac{h^3}{l} = 0, \tag{7·26}$$

whence $$\gamma = \frac{l}{3}\left\{-1\pm\left(-\frac{h^3}{l^3}\right)^{\frac{1}{2}}\right\}. \tag{7·27}$$

Since l is positive, γ will be complex only if h is positive; the corresponding vibration is damped.

Therefore the following result is obtained. There are no solutions for σ complex if h is negative, i.e. if the motion is unstable; but there may be solutions for complex values of σ, corresponding to damped vibrations, when the motion is stable.

It has to be borne in mind, however, that the characteristic equation will be complex in the latter case, so that certain additional relations will have to be satisfied by the two quantities h and l; since this case has no physical significance it will not be pursued further.

REFERENCES

Cornish, R. J. 1933 *Proc. Roy. Soc.* A, **140**, 227–240.
Fage, A. 1938 *Proc. Roy. Soc.* A, **165**, 513–517.
Faxen, H. 1928 *Ark. Mat. Astr. Fys.* **21** A, no. 26, 1–11.
Goldstein, S. 1937 *Proc. Camb. Phil. Soc.* **33**, 41–61.
Jeffreys, H. 1928 *Proc. Roy. Soc.* A, **118**, 195–208.
Lewis, J. W. 1927 *Proc. Roy. Soc.* A, **117**, 388–407.
Synge, J. L. 1938 *Proc. Roy. Soc.* A, **167**, 250–256.
Taylor, G. I. 1923 *Phil. Trans.* A, **223**, 289–343.
Taylor, G. I. 1936 *Proc. Roy. Soc.* A, **157**, 546–564.

ERRATUM

Vol. 186, page 396. First line of footnote:
For 'vorticity' *read* 'viscosity'.

The kinetics of the reactions of the steam-carbon system

By J. Gadsby, C. N. Hinshelwood, F.R.S., and K. W. Sykes

(*Received* 3 *January* 1946)

The kinetics of the various individual reactions which may occur in the steam-carbon system have been studied. The pressures of the separate gases have been varied in the range 10–760 mm. Essentially similar results have been obtained with coconut shell charcoal at 700° C and coal charcoal at 800° C.

The steam-carbon reaction, the primary product of which is carbon monoxide, is of fractional order with respect to steam and strongly retarded by hydrogen. The carbon dioxide-carbon reaction is of fractional order with respect to carbon dioxide and strongly retarded by carbon monoxide. The rates of both these reactions can be represented closely by an expression of the form

$$\text{rate} = \frac{k_1 p_1}{1 + k_2 p_2 + k_3 p_1},$$

where p_1 and p_2 are respectively the pressures of steam and hydrogen for the steam reaction, and of carbon dioxide and carbon monoxide for the carbon dioxide reaction. This kinetic scheme provides a consistent interpretation of the apparently conflicting results of previous work under a variety of conditions with many different types of carbon. Further experimental work, however, is necessary to elucidate without ambiguity the mechanisms of these reactions.

The water-gas reaction, $CO + H_2O = CO_2 + H_2$, takes place predominantly on the charcoal surface, and the approach to equilibrium has been studied from both sides. The forward reaction is of nearly the first order with respect to carbon monoxide and of fractional order with respect to steam; it is retarded by hydrogen and to a lesser extent by carbon dioxide. The reverse reaction is of fractional order with respect to both carbon dioxide and hydrogen, retarded by steam and unaffected by carbon monoxide. The kinetic expressions for the forward and reverse components of this heterogeneous reversible reaction combine to give the thermodynamically correct form of the equilibrium constant.

The reactions which may occur in the steam-carbon system are the following:

$$C + H_2O = CO + H_2, \tag{1}$$

$$C + 2H_2O = CO_2 + 2H_2, \tag{2}$$

$$CO + H_2O = CO_2 + H_2, \tag{3}$$

$$C + CO_2 = 2CO. \tag{4}$$

The kinetics of these reactions are not well understood in spite of the considerable amount of work, recently reviewed by Warner (1943), to which the industrial importance of the system has led. With two possible primary reactions, every product of which may enter into a secondary reaction, very carefully chosen conditions are necessary if the kinetic scheme is to be elucidated. This requirement has not often been met in investigations concerned largely with specific technical problems. As a first step in what is planned as a comprehensive study of all the separate reactions, a general survey of their kinetic relationships at a given temperature is described in this paper.

OUTLINE OF METHOD

A flow method is used in which a stream of gas is passed through a vertical column of granular charcoal kept at a known temperature. The partial pressures of all the gases present and the total rate of flow are controlled and varied separately, the total pressure being maintained at atmospheric by the use of nitrogen as an inactive carrier gas. The rate at which each separate product emerges from the charcoal column is measured. To elucidate the kinetics, the conditions are chosen so that one particular reaction predominates and interference by the other possible reactions is reduced to a minimum. This can be done by careful choice of the initial partial pressures and total rate of flow and by use of the fact that the steam-carbon and carbon dioxide-carbon reactions can largely be suppressed by hydrogen. The interpretation is further simplified by working with a large total rate of flow, so that only a small fraction of the reactant is consumed and consequently the partial pressure of the reactant and the total rate of flow are approximately uniform throughout the column.

Two types of active charcoal prepared by carbonization and steam activation, one from coconut shell and the other from briquetted coal, are used. Both are ordinary commercial samples containing small percentages of inorganic impurities which are known to affect the rate of reaction (Taylor & Neville 1921). This impure form of carbon is studied on account of its similarity to the material used industrially and its ready availability in large quantities. The general form of the results, however, is the same for the two different types of charcoal and is therefore probably of general significance and not dependent on the impurity content. All the results are obtained at 700° C for the nut charcoal and at 800° C for the coal charcoal, at which temperatures the rates of reaction for the two materials are comparable. The difference in reactivity of the two materials is probably due to their different impurity contents, since their surface areas are both of the order of 10^6 cm.^{2}/g. (Kipling, unpublished work).

Previous outgassing of the charcoal is unnecessary. The gas stream is passed over the charcoal until the rate of reaction reaches a steady value; then, with the total rate of flow constant, the partial pressure of a given gas is changed and the rate of reaction again measured when a new steady state has been reached. Fifteen minutes is sufficient for the steady state to be reached. To allow for the gradual change in the mass of charcoal present, the rate of reaction for a given composition of the gas stream is measured at least twice during each experiment, and by interpolation the results are adjusted to refer to 30 g. of the initial 35 g. of charcoal. Unless otherwise stated, the total rate of flow is 1200 c.c./min., measured at 20° C and 760 mm.

EXPERIMENTAL TECHNIQUE

35 g. of charcoal, sieved 8–10 B.S.S., occupy 10 cm. of a long vertical silica tube of diameter 3·1 cm. heated in an electric furnace *A* (figure 1). The temperature at the mid-point of the charcoal column is controlled to within ± 1° C by means of a thermocouple in conjunction with a potentiometer, a mirror galvanometer and a

relay operated by a photocell. It has since been discovered, however, that with an endothermic reaction and a flowing gas stream the temperature may not be uniform throughout the reaction zone. Subsequent measurements have shown that the variation of temperature at any other point may have been as much as ± 3° C and the range of temperature over the whole column as much as 15° C. Although the rates of these reactions change by about 2 % per ° C, it is thought that the general form of the results will not be appreciably affected. This has been confirmed by a more precise and more thorough investigation now in progress, in which not only the temperature at a point, but also the distribution of the temperature in space, is controlled automatically.

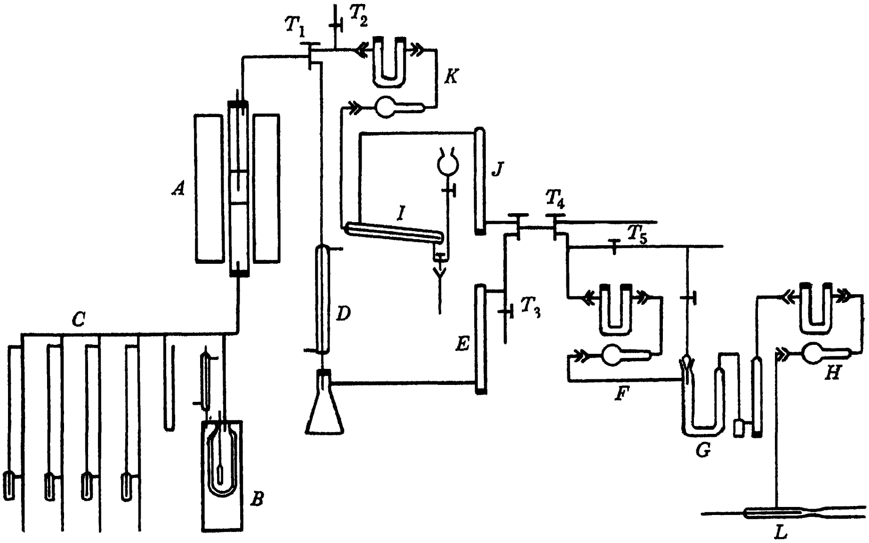


FIGURE 1. Diagram of apparatus.

The steam is generated in a calibrated boiler B, the rate of flow being controlled by the electric heating current. The other gases are obtained from cylinders and their rates of flow are measured by capillary tube flowmeters C, each calibrated for the particular gas used. All rates of flow are known to within ± 1 %. The only significant impurity, oxygen in the nitrogen, is removed by red-hot copper.

The rates of reaction are measured by diverting the gas stream for a known time through an analysis apparatus in which the products of reaction can be absorbed and weighed in detachable units. Carbon dioxide is absorbed by sofnolite (activated sodalime) followed by calcium chloride to take up the water formed. Carbon monoxide is first oxidized by air over hopcalite at 100° C and then absorbed by sofnolite. Each unit consists of a U-tube followed by a separate guard tube, fitted together and to the apparatus by ground-glass joints. For the steam-carbon reaction, the gas stream normally passes to waste through T_4, the steam having been removed by the condenser D and the calcium chloride tube E. To measure the rate of reaction the gas stream is diverted for a known time by T_4, the carbon dioxide being determined in F, the carbon monoxide oxidized in G and absorbed in H. Gas remaining in the dead space after T_4 has been returned to its normal position is swept into the absorption units by a stream of dry, carbon dioxide-free air from T_5. Any water formed by the oxidation of hydrogen by the hopcalite is removed by a calcium chloride tube between G and H. From the rates of flow of the gases, the stoichiometric equations and the estimations of carbon dioxide and carbon monoxide, complete information about the gas stream leaving the furnace is obtained. The same procedure is used for the carbon monoxide-steam reaction, the carbon dioxide alone being determined. For the carbon dioxide-carbon reaction, the gas stream normally passes to waste through D, E and T_3, the rate of formation of carbon monoxide being measured by the diversion of the gas stream by T_1 for a known time. The carbon dioxide is then removed by sodium hydroxide solution in I and sofnolite followed by calcium chloride in J, the carbon monoxide being determined in H after oxidation in G. The dead space is swept out by nitrogen from T_2. The procedure for the carbon dioxide-hydrogen reaction is similar. T_1 is again the controlling tap, the steam formed is determined in K, the carbon dioxide removed in I and J and the carbon monoxide determined in H. The extra pressure introduced by the diversion of the gas stream for analysis is balanced by the air pump L. The accuracy of the rate measurements is within 1 %, the time of absorption being several minutes and the increase in weight, determined to within 0·001 g., being not less than 0·100 g.

RESULTS

(1) *The steam-carbon reaction*

When steam is the chief constituent of the gas phase, comparable amounts of carbon monoxide and carbon dioxide are formed by the primary reaction of the steam with the carbon and by the water-gas reaction (equation (3)). The carbon dioxide-carbon reaction can be neglected, since under comparable conditions it is slower than the steam reaction by a factor of about three, and here the partial pressure of carbon dioxide is small compared with that of steam; it is, moreover, retarded by the hydrogen present. The sum of the rates of formation of carbon monoxide and carbon dioxide is therefore equal to the total rate of reaction of the steam with the carbon and is the quantity recorded here as the rate of reaction. No hydrocarbons could be detected with a Bone and Wheeler gas analysis apparatus.

The effect of steam. The apparent order of reaction with respect to steam is determined by varying the partial pressure, the rates of flow of steam and nitrogen being chosen so as to keep constant the total rate of flow. In figure 2 the rate of reaction is plotted against the initial partial pressure of steam. The reaction of both charcoals is of a similar fractional apparent order with respect to steam over the pressure range 100–760 mm. An apparent and not the true order of reaction is obtained since, although the fraction of steam reacting is small, the partial pressures of the products present are proportional to the rate of reaction. Since the reaction is retarded by hydrogen, the true order of reaction is greater than the apparent order under these conditions.

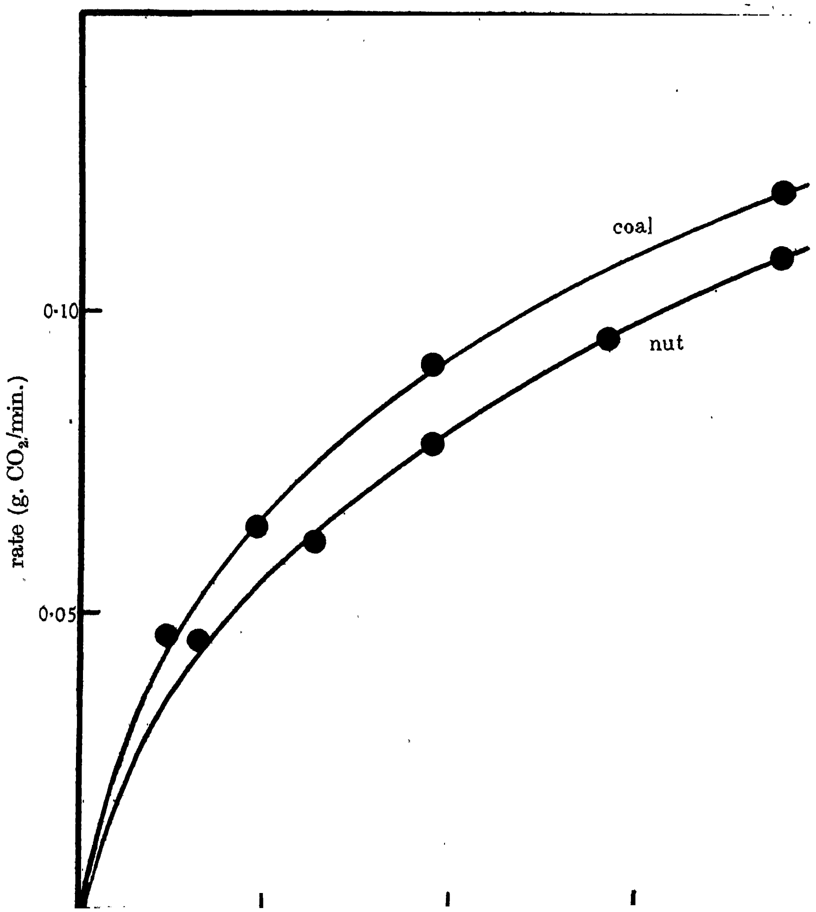


FIGURE 2. Steam-carbon reaction: apparent order with respect to steam.

The effect of hydrogen. Hydrogen at various small partial pressures is added, the rate of flow of nitrogen being adjusted so as to keep constant the partial pressure of steam and the total rate of flow. Figure 3, where the rate of reaction is plotted against the initial partial pressure of hydrogen, shows that the reaction of both types of charcoal is strongly retarded by hydrogen.

The effect of carbon monoxide. The results of analogous experiments with carbon monoxide are shown by the experimental points on the left-hand side of figure 4. The slight decrease in the rate of reaction is not a direct effect of the carbon monoxide introduced. Increase of the partial pressure of carbon monoxide increases the rate of formation of carbon dioxide and hydrogen from carbon monoxide and steam.

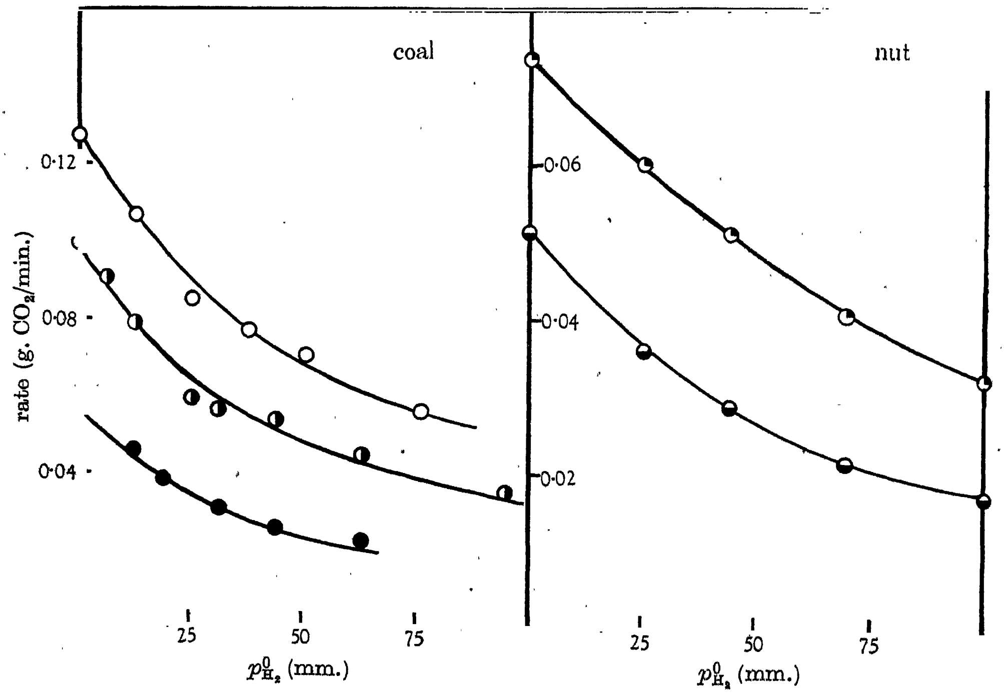


FIGURE 3. Steam-carbon reaction: retardation by hydrogen. Steam pressure: ○ 633 mm.; ◔ 507 mm.; ◑ 380 mm.; ◒ 253 mm.; ● 190 mm.

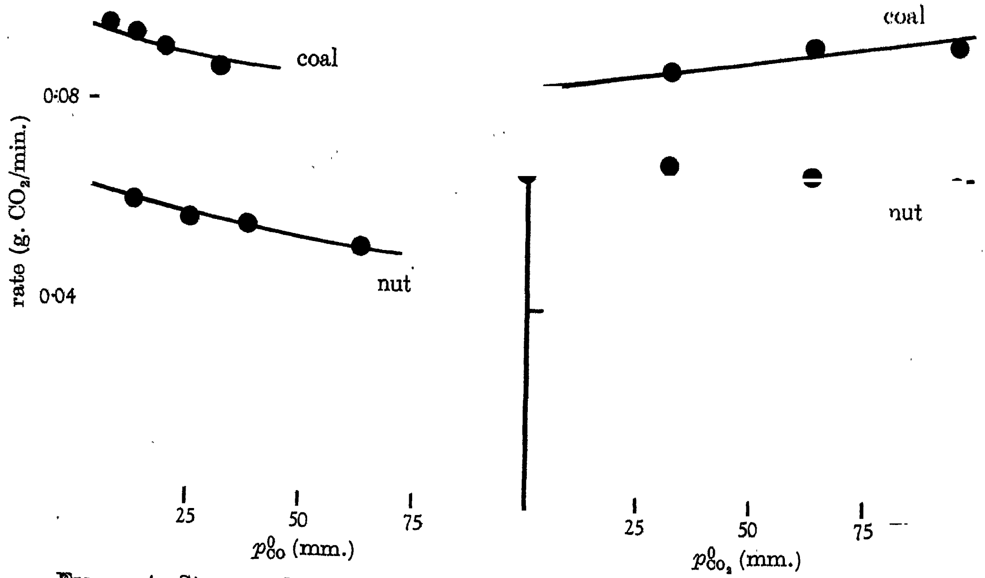


FIGURE 4. Steam-carbon reaction: effect of carbon monoxide and carbon dioxide.

The retarding effect of this increased partial pressure of hydrogen is shown by the curves themselves, which are calculated from the observed relation between the rate of reaction and the average partial pressure of hydrogen (figure 10). The agreement between the calculated curves and the experimental points shows that the decrease in the rate of reaction on the addition of carbon monoxide is accounted for entirely by the extra hydrogen formed in the water-gas reaction. For neither type of charcoal therefore is the rate of reaction appreciably affected by carbon monoxide.

The effect of carbon dioxide. The results of analogous experiments with carbon dioxide are shown on the right-hand side of figure 4. The rate of reaction of nut charcoal is not appreciably affected, and the slight increase in the rate of reaction of the coal charcoal can be accounted for by the carbon dioxide-carbon reaction.

The composition of the product. The results recorded in table 1 are qualitatively in good agreement with the usual assumption that the carbon dioxide is formed secondarily from carbon monoxide and steam and with the results obtained in this investigation for the kinetics of the secondary reaction. Carbon monoxide is far in excess of the equilibrium value, and since the water-gas reaction is of nearly the first order with respect to this gas, the fraction of it converted to dioxide will be approximately independent of its partial pressure. The effect on the composition of the product of increase in the partial pressure of steam is the resultant of two opposing influences: on the one hand the water-gas reaction is accelerated, though in less than direct proportion, and, on the other hand, it is retarded by the extra hydrogen formed in the primary reaction. The slight increase in the fraction of carbon dioxide formed for the nut charcoal and the decrease for the coal charcoal are consistent with the respective slight and considerable retardation by hydrogen of the water-gas reaction on those materials. The result of increasing the partial pressure of hydrogen is consistent with its effect on the water-gas reaction. The fraction of carbon dioxide calculated by numerical integration from the rate expressions for reactions 1 and 3 is smaller than the observed fraction by 11 % of the total oxides of carbon for the coal charcoal and by 24 % for the nut charcoal. On account of

TABLE 1. STEAM-CARBON REACTION: COMPOSITION OF PRODUCT

initial pressure (mm.)		fraction of CO_2 in oxides of carbon formed	
steam	hydrogen	coal	nut
127	0	0·304	0·559
253	0	0·263	0·599
380	0	0·245	0·633
570	0	0·231	0·671
760	0	0·225	0·694
380	0	0·246	0·654
380	12·7	0·237	0·641
380	25·3	0·193	0·646
380	44·3	0·172	0·636
380	63·2	0·164	0·633
380	88·5	0·156	0·623

experimental difficulty, however, this result is only a rough one. It gives an upper limit only for the possible primary carbon dioxide, which is thus shown to be small and may in fact be negligible.

(2) *The carbon dioxide-carbon reaction*

This reaction can be studied without interference by any of the others. Although it does not contribute appreciably to the steam-carbon system as a whole, its investigation is a necessary preliminary to that of the water-gas reaction. On account of their formal similarity, a comparison of the reactions of steam and of carbon dioxide with carbon might be expected to throw some light on their mechanism. The results given below are obtained with the coal charcoal; the kinetics of the reaction with nut charcoal, which are of the same general form, are being studied in a more accurate and more complete investigation now in progress.

The effect of carbon dioxide. The apparent order of reaction with respect to carbon dioxide is determined by varying the partial pressure, the rates of flow of carbon dioxide and nitrogen being chosen so as to keep constant the total rate of flow. The rate of reaction, defined for comparison with the steam reaction as one-half the rate of formation of carbon monoxide, is plotted against the initial partial pressure of carbon dioxide on the left-hand side of figure 5. The reaction is of a fractional apparent order with respect to carbon dioxide over the pressure range 50–500 mm., and comparison of figures 2 and 5 shows that it is slower than the steam reaction by

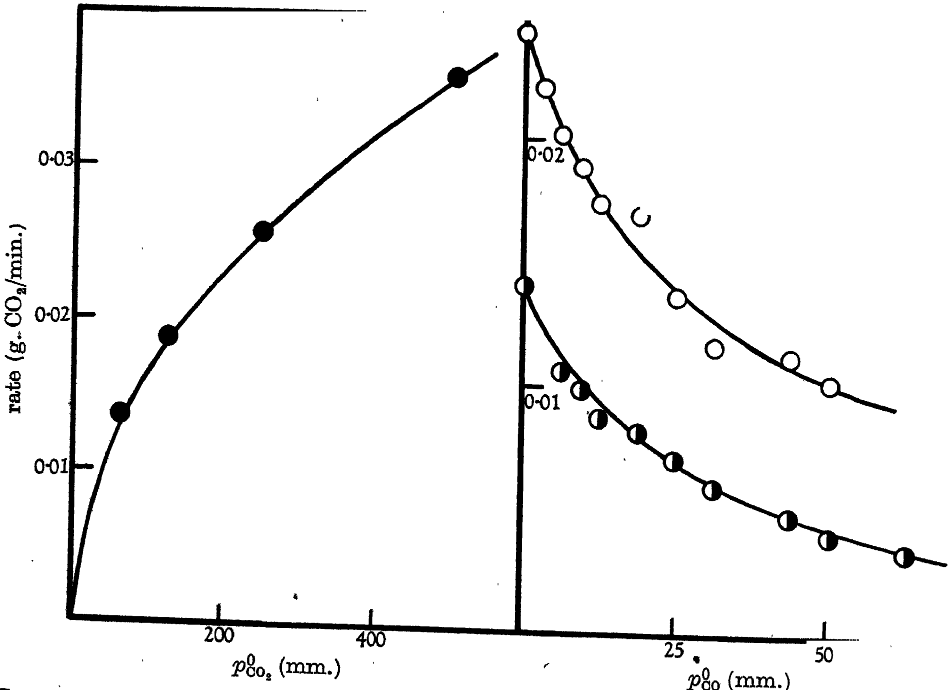


Figure 5. Carbon dioxide-carbon reaction: apparent order with respect to carbon dioxide and retardation by carbon monoxide. Carbon dioxide pressure: ○ 357 mm.; ◑ 126 mm.

a factor of about three. The true order of reaction is greater than the apparent order obtained under these conditions, since although the fraction of carbon dioxide reacting is small, the partial pressure of carbon monoxide which retards the reaction is proportional to the rate of reaction.

The effect of carbon monoxide. Various small partial pressures of carbon monoxide are added, the rate of flow of nitrogen being changed so as to keep constant the partial pressure of the reactant and the total rate of flow. The right-hand side of figure 5, where the rate of reaction is plotted against the initial partial pressure of carbon monoxide, shows that the reaction is strongly retarded by carbon monoxide. In all these measurements the rate of the reverse reaction, the disproportionation of carbon monoxide, is negligible compared with the rate of the forward reaction. This is shown by experiments in which carbon monoxide alone is passed over the charcoal, and by a calculation in which the equilibrium constant is assumed to be 8 atm. (Dent & Cobb 1929).

The effect of hydrogen. The reaction is also retarded by hydrogen as shown by table 2, which is discussed in connexion with the isolation of the water-gas reaction.

(3) *The water-gas reaction*

The equilibrium constant $[CO][H_2O]/[CO_2][H_2]$ being 0·63 at 700° C and 0·91 at 800° C (Bryant 1931), the forward and reverse reactions may take place at comparable rates, and both have been studied. Experiments with and without charcoal show that the rate of reaction either in the gas phase or on silica is negligible compared with the rate of reaction on the charcoal. Although the general form of the kinetics on the two charcoals is similar, the water-gas reaction depends more specifically on the charcoal than do the other reactions. On the coal charcoal in the absence of steam the rate of reaction is somewhat erratic and irreversibly reduced by previous exposure of the surface to hydrogen.

(*a*) *The carbon monoxide-steam reaction*

This reaction is studied by passing carbon monoxide and steam over the charcoal and measuring the rate of formation of carbon dioxide. The partial pressure of carbon monoxide is kept high to increase the rate of the water-gas reaction, and hydrogen is added to reduce the rate of the steam-carbon reaction. Under these conditions carbon dioxide is formed much more rapidly than without the added carbon monoxide, and can without any large error be taken as entirely due to the water-gas reaction. Further reaction of the carbon dioxide with the carbon is not appreciable in the presence of the hydrogen.

The effect of carbon monoxide and steam. The order of reaction with respect to carbon monoxide is determined by varying the partial pressure of the carbon monoxide, the partial pressures of steam and hydrogen and the total rate of flow being kept constant. Similarly, the order of reaction with respect to steam is determined by varying the partial pressure of steam, the partial pressures of carbon monoxide and hydrogen and the total rate of flow being kept constant. The results for the nut and coal charcoals are shown on the left-hand sides of figures 6 and 7 respectively, where the rate of reaction is plotted against the partial pressures of carbon monoxide and steam. A close approximation to the true order is obtained, since the rate of reaction is effectively the same throughout the charcoal column. The fraction of steam reacting is small, the increase in the partial pressure of hydrogen is small compared with its initial value, and the small fraction of carbon monoxide removed by the water-gas reaction is partly balanced by that formed in the steam-carbon reaction. With both charcoals the reaction is nearly of the first order with respect to carbon monoxide over the pressure range 50–300 mm., and of a low fractional order with respect to steam over the pressure range 100–600 mm.

The effect of hydrogen and carbon dioxide. Various small partial pressures of each product are added separately, the rate of flow of nitrogen being changed so as to keep constant the partial pressures of the other gases and the total rate of flow. The observed rate of reaction is corrected by the addition of the rate of the reverse reaction, calculated from the former, assuming

$$\frac{\text{rate of reverse reaction}}{\text{rate of forward reaction}} = K\frac{p_{CO_2}p_{H_2}}{p_{CO}p_{H_2O}}, \tag{5}$$

where K is the equilibrium constant as previously defined and the partial pressures are the averages of initial and final values; the correction is about 10 %. The upper curve on the right-hand side of figure 6 gives the rate of reaction against the average partial pressure of hydrogen and shows that the reaction on the nut charcoal is slightly retarded by hydrogen. The corresponding curve on the right-hand side of figure 7 shows that the reaction on the coal charcoal is more strongly retarded by hydrogen. The lower curve on the right-hand side of figure 6 shows that the reaction on the nut charcoal is slightly retarded by carbon dioxide.

(*b*) *The carbon dioxide-hydrogen reaction*

This reaction is studied by passing carbon dioxide and hydrogen over the charcoal and measuring the rates of formation of steam and carbon monoxide. If the water-gas reaction alone takes place, the rates of formation of steam and carbon monoxide will be equal. The carbon dioxide-carbon reaction increases the rate of formation of carbon monoxide, while the steam-carbon reaction increases the rate of formation of carbon monoxide and decreases the rate of formation of steam. The importance of these side reactions can therefore be determined by measuring the extent to which the rate of formation of carbon monoxide exceeds that of steam. In table 2 the rates of formation of steam and carbon monoxide are recorded for various initial partial pressures of hydrogen, the partial pressure of carbon dioxide and the total rate of flow being constant. With both charcoals the excess of carbon monoxide decreases rapidly as the partial pressure of hydrogen is increased and soon becomes negligible compared with the rate of the water-gas reaction; this means that not only the steam-carbon reaction, but also the carbon dioxide-carbon reaction is strongly retarded by hydrogen. It is therefore possible to choose the partial pressure of

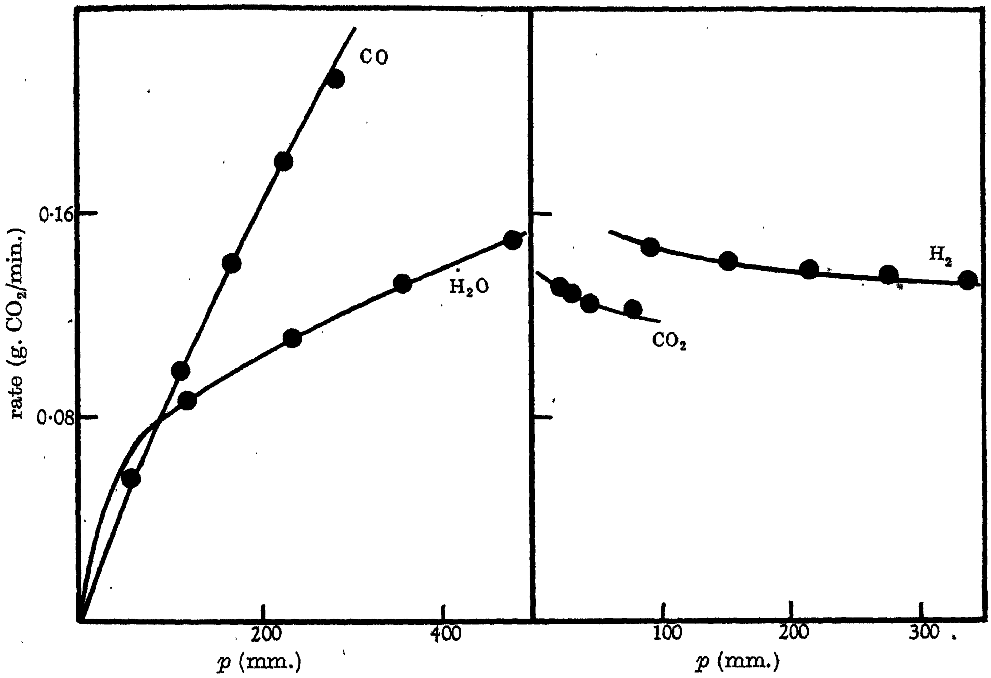


FIGURE 6. Carbon monoxide-steam reaction on nut charcoal: apparent order with respect to reactants and effect of products.

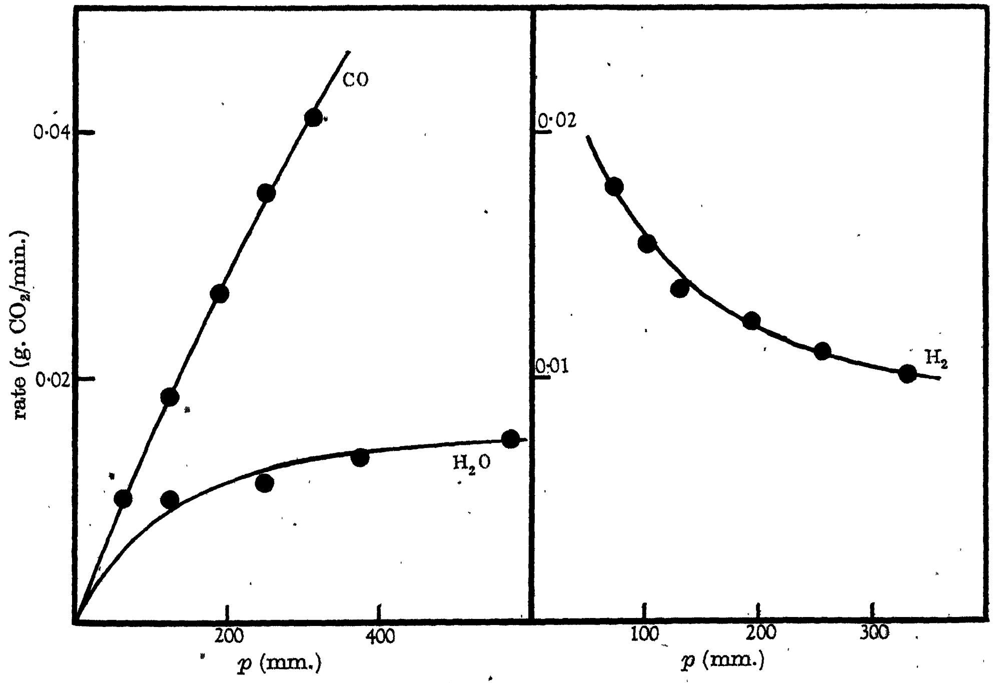


FIGURE 7. Carbon monoxide-steam reaction on coal charcoal: apparent order with respect to reactants and retardation by hydrogen.

hydrogen so that the rate of reaction of carbon dioxide with hydrogen can be measured without appreciable interference. For convenience, the rate of formation of steam rather than of carbon monoxide is measured.

Table 2. Carbon dioxide-hydrogen reaction: conditions for isolation

initial pressure (mm.)		rate of reaction (g. CO_2/min.)		
CO_2	H_2	to H_2O	to CO	excess CO
		nut charcoal		
253	0	—	0·0303	0·0303
253	32	0·0527	0·0655	0·0128
253	63	0·0920	0·0965	0·0045
253	95	0·1161	0·1186	0·0025
253	127	0·1350	0·1354	0·0004
		coal charcoal		
127	0	—	0·0450	0·0450
127	16	0·0289	0·0441	0·0152
127	32	0·0498	0·0579	0·0081
127	127	0·0953	0·0989	0·0036

The effect of carbon dioxide and hydrogen. Various partial pressures of each reactant are passed in turn over the charcoal in the presence of a constant partial pressure of the other reactant and at a constant total rate of flow. Before a series of measurements the surface of the charcoal is brought to a standard condition by reaction with steam. Although the fractions of carbon dioxide and hydrogen which react are small, apparent and not true orders are obtained since there is a retardation by steam. The results for the nut charcoal, which are reproducible, are shown on the left-hand side of figure 8; over the pressure range 25–500 mm. the apparent order of reaction is fractional with respect to each reactant, that with respect to hydrogen being the lower. The left-hand side of figure 9 shows that the results for the coal charcoal are very similar.

The effect of carbon monoxide and steam. Various small partial pressures of each product are added separately, the partial pressures of the other gases and the total rate of flow being kept constant. A small correction, up to about 10 %, is made for the reverse reaction by the method previously described (equation (5)). The right-hand side of figure 8, where the corrected rate of reaction is plotted against the average partial pressure of the product, shows that the reaction on the nut charcoal is unaffected by carbon monoxide but retarded by steam. The results for the coal charcoal are presented in a different way on account of the continual change in the rate of reaction. At constant partial pressures of carbon dioxide and hydrogen, the rate is measured successively with no added products, with 6 mm. of carbon monoxide added and with 6 mm. of steam added. The rates of reaction are plotted against time on the right-hand side of figure 9, the measurements with steam (filled circles) lying below the curve drawn through those with no added product (open circles) and those with carbon monoxide (half-filled circles). It is concluded that the reaction on the coal charcoal is unaffected by carbon monoxide but retarded by steam.

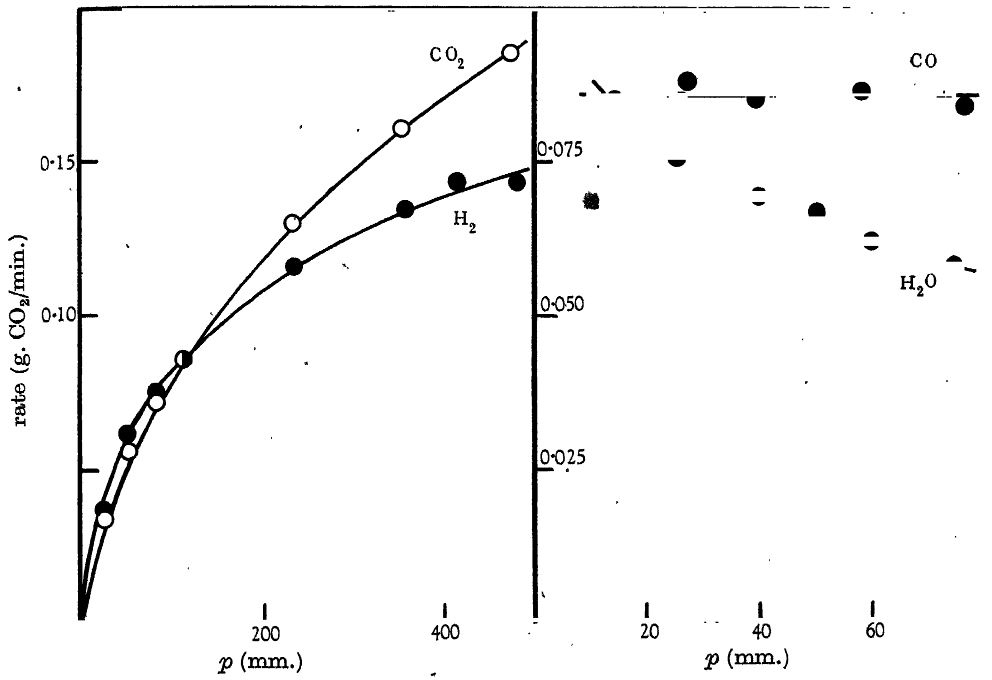


FIGURE 8. Carbon dioxide-hydrogen reaction on nut charcoal: apparent order with respect to reactants and effect of products.

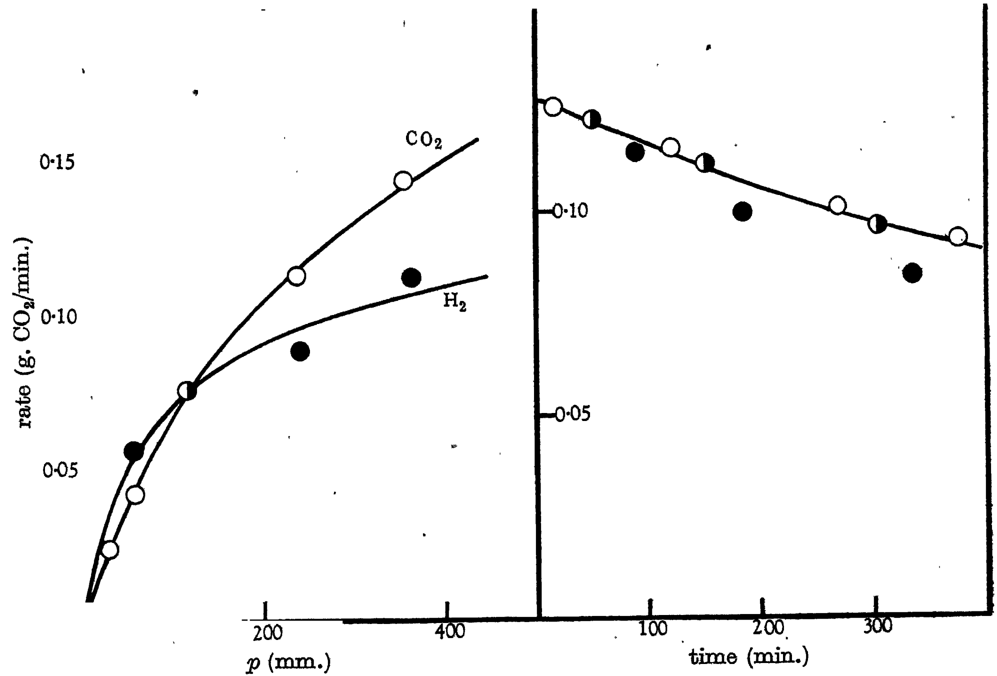


FIGURE 9. Carbon dioxide-hydrogen reaction on coal charcoal: apparent order with respect to reactants and effect of products.

DISCUSSION

(1) *The steam-carbon and carbon dioxide-carbon reactions*

The rate expression. The reaction of charcoal with steam is characterized by the fractional apparent order with respect to steam and the retardation by hydrogen, and the reaction with carbon dioxide by the fractional apparent order with respect to carbon dioxide and the retardation by carbon monoxide. Both reactions lead to the formation of carbon monoxide, and it will be shown that their rates can be represented by an expression of the form

$$\text{rate} = \frac{k_1 p_1}{1 + k_2 p_2 + k_3 p_1}, \tag{6}$$

where p_1 and p_2 are respectively the partial pressures of steam and hydrogen for the steam reaction or of carbon dioxide and carbon monoxide for the carbon dioxide reaction.

The relation between the expression under test and the observed rate of reaction can be deduced as follows. When a steady state has been reached, the net rate at which the reactant enters any volume element is equal to the rate of reaction in the volume element. Assuming that diffusion along the direction of flow can be neglected since the rates of diffusion are shown by rough calculation to be less than 5 % of the rates of flow, the condition for the steady state may be written

$$-\frac{V dp_1}{RT} = \left(\frac{k_1 p_1}{1 + k_2 p_2 + k_3 p_1}\right) \rho A \, dh, \tag{7}$$

where V is the total rate of flow at the temperature T of the reaction zone, dp_1 the increase in the partial pressure of the reactant as it passes through the volume element of length dh, the expression in brackets the rate of reaction in gram-molecules per unit time per unit mass of charcoal, A the cross-sectional area of the reaction tube and ρ the mass of charcoal per unit volume. Variations in the total rate of flow do not exceed 6 % and are neglected. The stoichiometric relation

$$p_2 = p_2^0 + g(p_1^0 - p_1) \tag{8}$$

is introduced, where the superscript zero denotes the initial value, $g = 1$ for the steam reaction and $g = 2$ for the carbon dioxide reaction. Integration over the reaction zone then gives

$$k = -(1 + g k_2 p_1^0 + k_2 p_2^0) \ln(1 - x) - (g k_2 - k_3) p_1^0 x, \tag{9}$$

where $x = (p_1^0 - p_1)/p_1^0$ is the fraction of the reactant which is consumed, k is defined by

$$k = \frac{k_1 RTm}{V}, \tag{10}$$

and $m = \rho A h$ is the total mass of charcoal present. Since $x < 0{\cdot}1$ the logarithm may be expanded with neglect of powers of x greater than the second and equation (9) becomes

$$k/x = 1 + k_3 p_1^0 + k_2 p_2^0 + \tfrac{1}{2} x (1 + g k_2 p_1^0 + k_2 p_2^0). \tag{11}$$

For the same reason the term in $\frac{1}{2}x$ in equation (11) will be relatively small; since $k_2 \gg 1$ (table 3) and $gp_1^0 > p_2^0$ it reduces to $xgk_2p_1^0/2$. Thus under the present conditions

$$\frac{1}{x} = \frac{1+k_3p_1^0}{k} + \frac{k_2}{k}\overline{p}_2, \tag{12}$$

where $\overline{p}_2$, the average partial pressure of the retarding product, is equal to the sum of the initial partial pressure and one-half the pressure increase in the reaction.

According to equation (12) the rate expression (equation (6)) can be tested by plotting $1/x$ against $\overline{p}_2$ for various values of p_1^0. The difficulty that some of the steam also reacts in the water-gas reaction can be minimized by assuming for the calculation of x that only carbon monoxide is formed and using the observed average partial pressure of hydrogen. Figure 10 shows that for each reaction the required series of

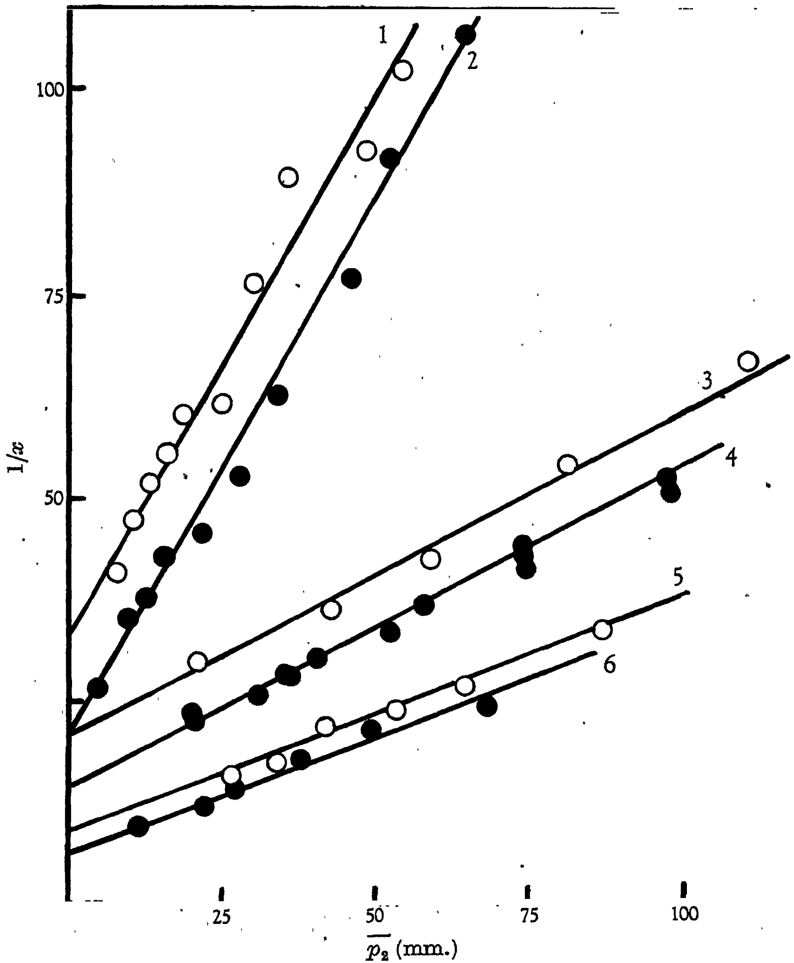


FIGURE 10. Steam-carbon and carbon dioxide-carbon reactions: test of rate expression. Pressure of reactant: CO_2—coal: 1, 357 mm.; 2, 127 mm. H_2O—nut (multiply ordinates by 0·75): 3, 506 mm.; 4, 380 mm. H_2O—coal: 5, 633 mm.; 6, 190 mm.

parallel straight lines with intercepts increasing with increase of p_1^0 is obtained within the limits of error. In figure 11 the intercept $1/x_0$ at $\overline{p_2} = 0$ is plotted against p_1^0. These intercepts are calculated by subtracting $k_2\overline{p_2}/k$ from values of $1/x$ obtained in experiments in which p_1^0 is varied without the addition of the retarding product: the values of k_2/k are given by the slopes of figure 10. The linear relation

$$\frac{1}{x_0} = \frac{1+k_3p_1^0}{k} \tag{13}$$

required by equation (12) is obeyed within the limits of error for each reaction. The rate expression therefore gives an adequate representation of the results. Values of the constants for the various reactions are given in table 3.

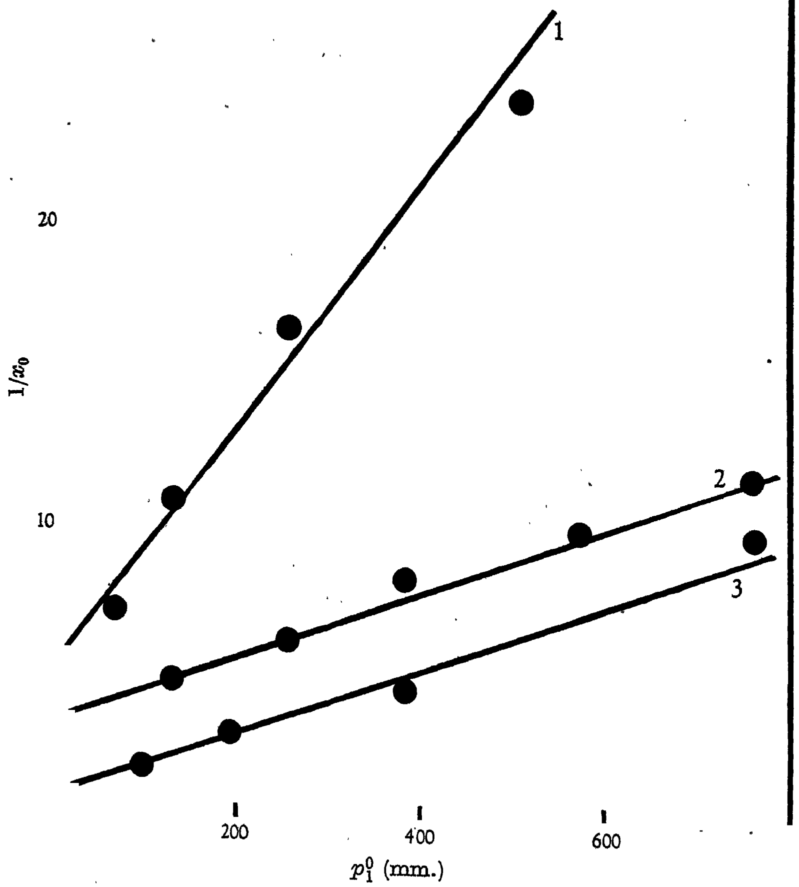


FIGURE 11. Steam-carbon and carbon dioxide-carbon reactions: test of rate expression. 1, CO_2—coal; 2, H_2O—nut; 3, H_2O—coal.

TABLE 3. STEAM-CARBON AND CARBON DIOXIDE-CARBON REACTIONS: VALUES OF CONSTANTS IN RATE EXPRESSION

reaction	k_1 (g.mol./min./atm./g.charcoal)	k_2 (atm.$^{-1}$)	k_3 (atm.$^{-1}$)
nut—H_2O, 700° C	$3{\cdot}6\times10^{-4}$	40	1·0
coal—H_2O, 800° C	$9{\cdot}0\times10^{-4}$	150	2·3
coal—CO_2, 800° C	$1{\cdot}4\times10^{-4}$	80	3·3

It is important to find out if the rate of reaction is determined partly by diffusion, since then the partial pressures at the surface will differ appreciably from those in the bulk of the gas phase and the correct rate expression for the surface reaction will not be obtained. An experimental criterion for the absence of control by diffusion is that the rate constants should not depend on the total rate of flow. With these reactions, however, there is the complicating factor that increase in the total rate of flow will reduce the partial pressure of the retarding product. Attention is therefore confined to the carbon dioxide reaction, where, in the absence of secondary reactions, the partial pressure of the retarding product can be calculated accurately. The experimental points in figure 12 show the rate of reaction as a function of the total rate of flow for three different initial pressures of carbon dioxide. The curves, which represent adequately the observed variation, are calculated from the rate expression (equation (6)) assuming that the constants k_1, k_2 and k_3 are independent of the total rate of flow. They are arranged to be equal to the observed rates at 1200 c.c./min. and are obtained by solving for x as a function of V in the equation

$$\frac{RTm}{xV}=\frac{1+k_3p_1^0}{k_1}+\frac{k_2}{k_1}xp_1^0, \tag{14}$$

which follows from equations (10) and (12). It therefore appears that the rate of reaction is not controlled by diffusion outside the charcoal granules, and in view of the high temperature coefficient, it is unlikely that the general form of the results is affected by diffusion inside the granules. The increase of the rate of reaction to a limiting value as carbon dioxide or steam is passed at an increasing rate of flow over graphite or coke has been observed before, but the whole effect has been attributed to diffusion without the possibility of specific retardations being considered (Mayers 1934*a*, *b*, 1939).

The steam-carbon reaction has previously been reported to be of negative order, of zero order, of the first order and of the second order with respect to steam. Although this lack of agreement may result in part from the different types of carbon and conditions of reaction used, it is largely due to the unjustified assumption that the rate of reaction is independent of the pressure of hydrogen. Key & Cobb (1930), working with various cokes at 1000° C, found an apparent order of between the first and the second by varying the column length, and, having observed qualitatively that the reaction was retarded by hydrogen, assumed that the true order was the first. Using the results for the rate of flow variation obtained by Brewer & Reyerson

(1934) with 'lignite char' in the temperature range 700–1200° C, Scott (1941) found that the apparent order with respect to steam was close to the second, and, without mentioning the possibility that the reaction was retarded, assumed that the true order was the second. The apparent order of reaction determined by this method, however, is greater than the true order, since increase in the pressure of steam is accompanied by decrease in the pressure of hydrogen.

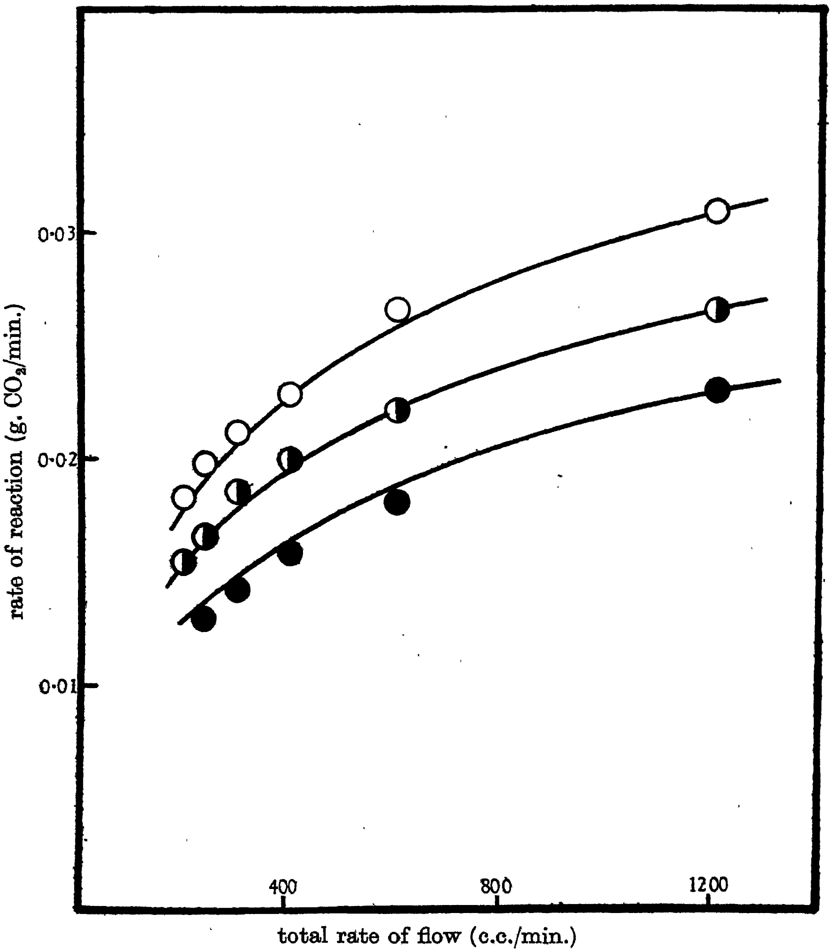


FIGURE 12. Carbon dioxide-carbon reaction: variation of total rate of flow. Carbon dioxide pressure: ○ 506 mm.; ◑ 380 mm.; ● 253 mm.

Another two investigations gave apparently conflicting results. In the first (Haslam, Hitchcock & Rudow 1923) the volume of steam passing in unit time (total rate of flow) was kept constant while the pressure was varied, and the order of reaction of electrode carbon in the temperature range 650–1200° C was found to be slightly less than the first, the experimental results being similar to those of figure 2. In the second (Thiele & Haslam 1927) the mass of steam passing in unit time was kept constant and the reactions of the same electrode carbon, of active charcoal and of graphite were found to be of zero order, while the rate of reaction of retort carbon decreased with increase of steam pressure (negative order). However, in the second investigation two opposing factors were operating, since increase in the pressure of steam was necessarily accompanied by a proportional decrease in the total rate of flow and thus an increase in the pressure of hydrogen. The rate of reaction being almost inversely proportional to the pressure of hydrogen, a reaction of the first order would appear to be of zero order and a reaction or fractional order to be of negative order under those conditions.

The most recent work on the carbon dioxide reaction is that of Frank-Kamenetzky & Semechkova (1940) with a static system. The reactions of sugar charcoal, with and without the addition of iron and aluminium, and of coke were interpreted by means of an expression of the same form as equation (6). The effect of carbon monoxide, however, was not measured directly.

The rate expression proposed in the present paper is thus by no means incompatible with previous experimental results and provides a consistent interpretation of work under a variety of conditions with many different types of carbon.

Reaction mechanism. There are several different types of reaction mechanism which lead to the observed rate expression. In the following discussion the reactant, steam or carbon dioxide, is represented by RO, and the retarding product, hydrogen or carbon monoxide, by R, since for each reaction it is the molecule obtained by removing an atom of oxygen from the reactant; molecules chemisorbed on the carbon surface are represented by symbols enclosed in brackets.

The simplest assumptions are that both reactant and retarding product are adsorbed as such on the carbon surface, and that the rate of reaction is proportional to the fraction of the surface covered by the reactant:

$$RO \underset{k_2}{\overset{k_1}{\rightleftharpoons}} (RO), \tag{15}$$

$$R \underset{k_4}{\overset{k_3}{\rightleftharpoons}} (R), \tag{16}$$

$$\mathrm{C} + (RO) \xrightarrow[k_5]{} \mathrm{CO} + R. \tag{17}$$

The conditions for a steady state on a surface of equivalent and independent sites, each of which can be occupied by one molecule of reactant or retarding product, are

$$k_1 p_1(1-\theta_1-\theta_2) = (k_2+k_5)\,\theta_1, \tag{18}$$

$$k_3 p_2(1-\theta_1-\theta_2) = k_4\theta_2, \tag{19}$$

where θ_1 and θ_2 are the fractions of the surface occupied by reactant and retarding product respectively. The rate of reaction is given by

$$\text{rate} = \frac{\dfrac{k_1 k_5 p_1}{k_2+k_5}}{1+\dfrac{k_3 p_2}{k_4}+\dfrac{k_1 p_1}{k_2+k_5}}, \tag{20}$$

which is of the same form as the observed rate expression (equation (6)). This mechanism can also account for the rather striking fact that in each case the reaction of RO with carbon is retarded by R. If the adsorption of the reactant requires its dissociation so that (RO) represents an atom of oxygen and the retarding product adsorbed on neighbouring sites, the reaction of RO will necessarily take place on sites which can also adsorb R.

Another scheme of general applicability is suggested by the mechanism which Frank-Kamenetzky & Semechkova (1940) proposed for the carbon dioxide reaction

$$RO \underset{k_2}{\overset{k_1}{\rightleftharpoons}} (\mathrm{O}) + R, \tag{21}$$

$$R \underset{k_4}{\overset{k_3}{\rightleftharpoons}} (R), \tag{22}$$

$$\mathrm{C} + (\mathrm{O}) \xrightarrow[k_5]{} \mathrm{CO}. \tag{23}$$

The reactant is not adsorbed as such, but reacts with the carbon to give an atom of oxygen which remains on the surface and the rest of the molecule which passes into the gas phase. The retarding product may decrease the rate of reaction either by adsorption or by reaction with adsorbed oxygen atoms according to the reverse component of equation (21). If one or the other of these two distinct mechanisms of retardation predominates, a rate expression of the desired form is obtained. The conditions for a steady state on the surface are

$$k_1 p_1 (1 - \theta_1 - \theta_2) = (k_5 + k_2 p_2)\,\theta_1, \tag{24}$$

$$k_3 p_2 (1 - \theta_1 - \theta_2) = k_4 \theta_2, \tag{25}$$

where θ_1 and θ_2 are now the fractions of the surface occupied by oxygen atoms and retarding product respectively. If the retardation is largely due to the adsorption of the retarding product and k_2 can be neglected, the rate of reaction is given by

$$\text{rate} = \frac{k_1 p_1}{1 + \dfrac{k_3 p_2}{k_4} + \dfrac{k_1 p_1}{k_5}}. \tag{26}$$

If the retardation is largely due to reaction with adsorbed oxygen atoms (k_2) and the adsorption of the retarding product (k_3) can be neglected, the rate of reaction is given by

$$\text{rate} = \frac{k_1 p_1}{1 + \dfrac{k_2 p_2}{k_5} + \dfrac{k_1 p_1}{k_5}}. \tag{27}$$

In order to decide between these various possibilities, it is necessary to study the adsorption of the various gases and the temperature dependence of the constants of the rate expression. It appears that carbon dioxide is not adsorbed as such by carbon at high temperatures, but reacts with the surface according to equation (21) (Langmuir 1915; Broom & Travers 1932; Frank-Kamenetzky & Semechkova 1940); it is therefore probable that the mechanism of the carbon dioxide reaction is that either of Frank-Kamenetzky & Semechkova, or of equation (26), or of equation (27).

For steam the evidence is less plentiful, but the results of Muller & Cobb (1940) suggest that the whole molecule is adsorbed by carbon at high temperatures. Since hydrogen is also adsorbed by carbon at high temperatures (Barrer 1935; Muller & Cobb 1940), it is probable that the mechanism of the steam reaction is that of equations (15)–(17). More experimental work, however, must be carried out before the detailed mechanisms of these reactions can be elucidated without ambiguity.

(2) *The water-gas reaction*

So far as we are aware, no experimental data on the kinetics of the water-gas reaction catalysed by charcoal have previously been reported. From the results for the nut and coal charcoals taken together, it appears that the reaction between carbon monoxide and steam is of the first order with respect to carbon monoxide, of fractional order with respect to steam, retarded by hydrogen and to a lesser extent by carbon dioxide, and that the reaction between carbon dioxide and hydrogen is of fractional order with respect to carbon dioxide and hydrogen, retarded by steam and unaffected by carbon monoxide. Although the results leave in no doubt the substantial correctness of this general statement, it has not been possible to determine with certainty the exact form of the rate expressions. No simple expression is applicable and the lack of reproducibility, possibly occasioned by changes in catalytically active impurities, makes it difficult to distinguish between various types of more complicated expressions. The detailed mechanism of the reaction will therefore not be discussed and attention will be confined to a comparison of the more general features of the forward and reverse components of this reaction.

Comparison of forward and reverse reactions. The kinetics of the two components of a reversible reaction must be related so as to yield at equilibrium the correct equilibrium constant. For a number of homogeneous reactions this condition is satisfied in the simplest possible way, the rates of reaction being proportional to the stoichiometric product, and the rate constants being independent of whether or not the system is at equilibrium (for a summary, see Glasstone 1940). For heterogeneous reactions, however, the rate expressions are often complicated, and it is not easy to see at once any simple way in which the equilibrium condition can be satisfied. However, the assumption that the rate expressions are the same whether or not the system is in equilibrium leads to a simple condition with which the results for the water-gas reaction are in complete qualitative agreement.

Consider any chemical reaction $A+B \rightleftharpoons C+D$. Let the rate of the forward reaction be represented by $k_1 f_1(a,b,c,d)$ and that of the reverse reaction by $k_2 f_2(a,b,c,d)$, where a, b, c and d are the concentrations of A, B, C and D respectively. Now define the functions $F_1(a,b,c,d)$ and $F_2(a,b,c,d)$ by the relations

$$f_1(a,b,c,d) = \frac{ab}{F_1(a,b,c,d)}, \tag{28}$$

$$f_2(a,b,c,d) = \frac{cd}{F_2(a,b,c,d)}. \tag{29}$$

Since the rates of the forward and reverse reactions are equal at equilibrium, the following may be written for the equilibrium constant:

$$K = \frac{cd}{ab} = \frac{k_1 F_2(a, b, c, d)}{k_2 F_1(a, b, c, d)}. \tag{30}$$

In order that K shall be constant under all conditions, neglecting deviations from ideal behaviour, then

$$F_1(a, b, c, d) = F_2(a, b, c, d) \tag{31}$$

for all values of the concentrations. Thus if the rate expressions are written in the form of the stoichiometric product divided by some function of the concentrations, the denominator will be exactly the same for both components of the reaction. If the forward reaction is of the first order with respect to A, there will be no term in a in the denominator and the reverse reaction will be unaffected by A. If the forward reaction is of a low fractional order with respect to B, there will be a large term in b in the denominator and the reverse reaction will be retarded by B. The evidence for the reaction on the nut charcoal is as follows. The left-hand side of figure 6 shows that the forward reaction is of nearly the first order with respect to carbon monoxide and of a low fractional order with respect to steam; the right-hand side of figure 8 shows that the reverse reaction is unaffected by carbon monoxide and retarded by steam. The left-hand side of figure 8 shows that the reverse reaction is of fractional apparent order with respect to both carbon dioxide and hydrogen; the right-hand side of figure 6 shows that the forward reaction is retarded by both carbon dioxide and hydrogen. Similar evidence for the reaction on the coal charcoal is provided by the corresponding portions of figures 7 and 9.

The physical interpretation of equation (31) for a heterogeneous reaction can readily be given. The effects of a gas on the forward and reverse reactions are necessarily related, since they are both determined by the extent to which it is adsorbed. For example, if it is only slightly adsorbed, the reaction in which it takes part will be of the first order with respect to it and it will not retard the reverse reaction, since it does not occupy an appreciable fraction of the surface. This aspect has been discussed before (Hinshelwood 1933), but, so far as we are aware, the experimental evidence has not previously been obtained.

This work forms part of the programme of the Fuel Research Board of the Department of Scientific and Industrial Research, and this paper is published by permission of the Department. One of the authors (J.G.) was a member of the Department's staff. The illustrations are Crown Copyright and are published by permission of the Controller of His Majesty's Stationery Office.

REFERENCES

Barrer 1935 *Proc. Roy. Soc.* A, **149**, 253.
Brewer & Reyerson 1934 *Industr. Engng Chem.* **26**, 734.
Broom & Travers 1932 *Proc. Roy. Soc.* A, **135**, 512.
Bryant 1931 *Industr. Engng Chem.* **23**, 1019.

Dent & Cobb 1929 *J. Chem. Soc.* p. 1903.
Frank-Kamenetzky & Semechkova 1940 *Acta Physicochim. U.R.S.S.* **12**, 879.
Glasstone 1940 *Textbook of physical chemistry*, p. 1053. Macmillan.
Haslam, Hitchcock & Rudow 1923 *Industr. Engng Chem.* **15**, 115.
Hinshelwood 1933 *Kinetics of chemical change in gaseous systems*, 3rd ed. p. 346.
Key & Cobb 1930 *J. Soc. Chem. Ind., Lond.* **49**, 439T.
Langmuir 1915 *J. Amer. Chem. Soc.* **37**, 1139.
Mayers 1934*a* *J. Amer. Chem. Soc.* **56**, 70.
Mayers 1934*b* *J. Amer. Chem. Soc.* **56**, 1879.
Mayers 1939 *J. Amer. Chem. Soc.* **61**, 2053.
Muller & Cobb 1940 *J. Chem. Soc.* p. 177.
Scott 1941 *Industr. Engng Chem.* **33**, 1279.
Taylor & Neville 1921 *J. Amer. Chem. Soc.* **43**, 2055.
Thiele & Haslam 1927 *Industr. Engng Chem.* **19**, 882.
Warner 1943 *J. Amer. Chem. Soc.* **65**, 1447.

On lattice points in n-dimensional star bodies

I. Existence theorems

By K. Mahler, *Manchester*

(*Communicated by L. J. Mordell, F.R.S.—Received* 27 *April* 1945)

Let $F(X) = F(x_1, \ldots, x_n)$ be a continuous non-negative function of X satisfying $F(tX) = |t|F(X)$ for all real numbers t. The set K in n-dimensional Euclidean space R_n defined by $F(X) \leqslant 1$ is called a star body. The author studies the lattices Λ in R_n which are of minimum determinant and have no point except $(0, \ldots, 0)$ inside K. He investigates how many points of such lattices lie on, or near to, the boundary of K, and considers in detail the case when K admits an infinite group of linear transformations into itself.

INTRODUCTION

Let K be an arbitrary bounded or unbounded point set in the n-dimensional Euclidean space R_n of all points

$$X = (x_1, x_2, \ldots, x_n) \qquad (x_1, x_2, \ldots, x_n \text{ real numbers}).$$

A point lattice Λ,

$$x_h = \sum_{k=1}^{n} a_{hk} u_k \quad (h = 1, 2, \ldots, n,\ u_1, u_2, \ldots, u_n = 0, \pm 1, \pm 2, \ldots),$$

in R_n of determinant

$$d(\Lambda) = \left| \, |a_{hk}|_{h,k=1,2,\ldots,n} \right|$$

is called K-admissible if no point P of Λ, except possibly the origin $O = (0, 0, \ldots, 0)$, is an inner point of K. (P is an inner point of K if there is an n-dimensional sphere with centre at P and contained in K.) The minimum determinant $\Delta(K)$ of K is defined as the lower bound of $d(\Lambda)$ extended over all K-admissible lattices. This function $\Delta(K)$ depends on K in a very complicated way and is, in general, not a continuous function of K. A K-admissible lattice Λ such that $d(\Lambda) = \Delta(K)$ is called a critical lattice of K; such critical lattices exist, for instance, if K contains O as an inner point and has at least one admissible lattice.

Minkowski proved in his classical theorem that if K is a convex body with centre at O, then

$$2^n \Delta(K) \geqslant V(K),$$

where $V(K)$ is the volume of K. He further gave a finite algorism for obtaining $\Delta(K)$ and the critical lattices of K if K is such a convex body and $n = 2$ or $n = 3$, or if K is of a certain type with $n = 4$ (Minkowski 1907, 1911).

Minkowski also considered another more general class of point sets, the star bodies (*Strahlenkörper*). These are point sets defined by an inequality

$$F(X) \leqslant 1,$$

where $F(X) = F(x_1, \ldots, x_n)$ is a continuous function of X such that

$$F(X) \geqslant 0 \quad \text{for all points } X,$$

$$F(tx_1, \ldots, tx_n) = |t| \, F(x_1, \ldots, x_n) \quad \text{for real } t.$$

The functional equation implies that K is symmetrical in O. This restriction is not made by Minkowski, but is in no way essential. He found (1911) for such point sets that

$$2\zeta(n) \Delta(K) \leqslant V(K),$$

but his proof was never published. Recently, Hlawka (1943) gave a very ingenious proof based on the theory of multiple integrals, and I found a geometrical proof (Mahler 1944) for a slightly less exact inequality.*

New progress was made in the years from 1938 onwards when important special examples of star bodies in two or three dimensions were investigated by Davenport (1938, 1939 and 1944) and Mordell (1942, 1943, 1944, and the general method 1945). In 1941 Mordell discovered a method for dealing with a certain important class of such problems. This work led me to ask myself whether Minkowski's method of evaluating $\Delta(K)$ when K is convex (Minkowski 1907, 1911) could be extended to arbitrary bounded star bodies. I succeeded in answering this question in the affirmative, and found an algorism for the evaluation of $\Delta(K)$ if K is two-dimensional and bounded; and I applied this method to a few special cases.

In the present paper, the aim is not to consider further special examples of star bodies, but rather to lay the foundations of a general theory of bounded or unbounded n-dimensional star bodies and their critical lattices.

In this first part, I begin by proving that if the star body K,

$$F(X) \leqslant 1,$$

---

* *Addition, May* 1946. A beautiful new proof of the Minkowski-Hlawka theorem was recently given by C. L. Siegel, *Ann. Math.* **46** (1945), 340–347.

has at least one admissible lattice, then K also admits at least one critical lattice. The points of such a critical lattice Λ on, or in the neighbourhood of, the boundary C of K are next studied. If K is bounded, then at least $2n$ points of Λ lie on C, as is almost obvious; an example is constructed in which this lower bound is attained. If K is not bounded, then Λ need not have a single point on C, as is also proved by means of an example. It is then easily proved that to every $\epsilon > 0$ there is at least one point P of Λ such that

$$1 \leqslant F(P) < 1+\epsilon;$$

however, it remains an open question whether there are always n independent points of Λ with this property.

From § 14 onwards, unbounded star bodies are considered with an infinite group Γ of linear transformations into themselves; many of the most interesting lattice-point problems are of this type. Three different assumptions about Γ are made and applied to the study of the critical lattices. Then three general classes of star bodies are found with the following three properties respectively: (*a*) At least one critical lattice of K has a point on C (theorem 21). (*b*) For every $\epsilon > 0$, every critical lattice Λ of K contains an infinity of points P satisfying

$$1 \leqslant F(P) < 1+\epsilon$$

(theorem 23). (*c*) For every $\epsilon > 0$, every critical lattice Λ of K contains n independent points $P_1, \ldots, P_n$ satisfying

$$1 \leqslant F(P_g) < 1+\epsilon \quad (g = 1, 2, \ldots, n)$$

(theorem 25). The simplest example of an n-dimensional star body with all three properties (*a*), (*b*), (*c*) is that defined by the inequality

$$|\, x_1 x_2 \ldots x_n \,| \leqslant 1.$$

In the second part of this paper which is appearing in the *Proc. Royal Acad. Amsterdam*, I intend to study certain types of star bodies K according as they contain, or do not contain, smaller star bodies K' such that

$$\Delta(K') = \Delta(K).$$

1. NOTATION

The following notation is used in this paper:

If $x_1, x_2, \ldots, x_n$ $(n \geqslant 2)$ are real numbers, then

$$X = (x_1, x_2, \ldots, x_n) \tag{1·1}$$

is the point in n-dimensional Euclidean space R_n with rectangular co-ordinates $x_1, x_2, \ldots, x_n$. The non-negative number

$$|\, X \,| = +(x_1^2 + x_2^2 + \ldots + x_n^2)^{\frac{1}{2}} \tag{1·2}$$

is called the distance of X from the origin $O = (0, 0, \ldots, 0)$. If

$$X_1 = (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}), \ldots, X_r = (x_1^{(r)}, x_2^{(r)}, \ldots, x_n^{(r)}) \tag{1·3}$$

are any points in R_n, and $\lambda_1, \ldots, \lambda_r$ are any real numbers, then

$$\lambda_1 X_1 + \ldots + \lambda_r X_r$$

is written for the point

$$(\lambda_1 x_1^{(1)} + \ldots + \lambda_r x_1^{(r)}, \lambda_1 x_2^{(1)} + \ldots + \lambda_r x_2^{(r)}, \ldots, \lambda_1 x_n^{(1)} + \ldots + \lambda_r x_n^{(r)}).$$

The determinant of n points

$$X_1 = (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}), \ldots, X_n = (x_1^{(n)}, x_2^{(n)}, \ldots, x_n^{(n)}) \tag{1·4}$$

is denoted by
$$\{X_1, X_2, \ldots, X_n\} = \begin{vmatrix} x_1^{(1)} & x_2^{(1)} & \ldots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \ldots & x_n^{(2)} \\ \vdots & \vdots & & \vdots \\ x_1^{(n)} & x_2^{(n)} & \ldots & x_n^{(n)} \end{vmatrix}. \tag{1·5}$$

The points are called independent, if this determinant does not vanish.

The set Λ of all points

$$X = u_1 X_1 + \ldots + u_n X_n, \quad \text{where} \quad u_1, \ldots, u_n = 0, \pm 1, \pm 2, \pm 3, \ldots,$$

is called a lattice if its determinant

$$d(\Lambda) = |\{X_1, X_2, \ldots, X_n\}| \tag{1·6}$$

is not zero; then $X_1, X_2, \ldots, X_n$ are said to form a basis of Λ. Any n points $Y_1, Y_2, \ldots, Y_n$ of Λ form a basis of this lattice if and only if

$$\{Y_1, Y_2, \ldots, Y_n\} = \pm d(\Lambda). \tag{1·7}$$

If $P, Q, R, \ldots$ are points of Λ, then $\Lambda - [P, Q, R, \ldots]$ denotes the set of all points of Λ different from $P, Q, R, \ldots$.

2. THE REDUCED BASIS OF A LATTICE

THEOREM 1. *There exists a constant $\gamma_n > 0$ depending only on the dimension n of R_n, with the following property: every lattice Λ in R_n has a reduced basis, i.e. a basis $Y_1, Y_2, \ldots, Y_n$ for which*

$$|Y_1|\,|Y_2| \ldots |Y_n| \leqslant \gamma_n d(\Lambda). \tag{2·1}$$

Proof. Let
$$X_1 = (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}), \ldots, X_n = (x_1^{(n)}, x_2^{(n)}, \ldots, x_n^{(n)}) \tag{2·2}$$

be any basis of Λ. Then

$$\Phi(u_1, \ldots, u_n) = \sum_{g=1}^{n} (x_g^{(1)} u_1 + \ldots + x_g^{(n)} u_n)^2 = |u_1 X_1 + \ldots + u_n X_n|^2 \tag{2·3}$$

is a positive definite quadratic form of discriminant

$$d(\Lambda)^2 = \{X_1, X_2, \ldots, X_n\}^2. \tag{2·4}$$

There exists a linear unimodular substitution

$$u_g = \sum_{h=1}^{n} a_{gh} v_h, \quad \text{where} \quad g = 1, 2, \ldots, n, \tag{2·5}$$

with integral coefficients by which Φ is changed into a new form

$$\Phi(u_1, \ldots, u_n) = \Psi(v_1, \ldots, v_n) = \sum_{g=1}^{n} (y_g^{(1)} v_1 + \ldots + y_g^{(n)} v_n)^2, \tag{2·6}$$

which is reduced according to Minkowski (1911). Hence by his theorem

$$\Psi(1, 0, \ldots, 0)\, \Psi(0, 1, \ldots, 0) \ldots \Psi(0, 0, \ldots, 1) \leqslant \gamma_n^2 d(\Lambda)^2, \tag{2·7}$$

where $\gamma_n > 0$ depends only on n. The n points

$$Y_1 = (y_1^{(1)}, y_2^{(1)}, \ldots, y_n^{(1)}), \ldots, Y_n = (y_1^{(n)}, y_2^{(n)}, \ldots, y_n^{(n)}) \tag{2·8}$$

form a basis of Λ since

$$\{Y_1, \ldots, Y_n\} = \{X_1, \ldots, X_n\} = \pm d(\Lambda). \tag{2·9}$$

Moreover,

$$\Psi(1, 0, \ldots, 0) = |Y_1|^2, \; \Psi(0, 1, \ldots, 0) = |Y_2|^2, \; \ldots, \; \Psi(0, 0, \ldots, 1) = |Y_n|^2, \tag{2·10}$$

whence the assertion.

Theorem I may also be proved by the reduction method of Hermite (1905), which has the advantage that the proof of the product formula for the Ψ's is of an elementary character.

3. THE CONVERGENCE THEOREM

DEFINITION 1. *An infinite sequence of lattices*

$$\Lambda_1, \Lambda_2, \Lambda_3, \ldots$$

is called bounded, if there exist two positive numbers c_1, c_2 *such that*

$$d(\Lambda_r) \leqslant c_1 \quad \textit{for} \quad r = 1, 2, 3, \ldots; \tag{3·1}$$

$$|X| \geqslant c_2 \quad \textit{for all points } X \neq O \textit{ of } \Lambda_r, \quad \textit{when} \quad r = 1, 2, 3, \ldots. \tag{3·2}$$

DEFINITION 2. *An infinite sequence of lattices*

$$\Lambda_1, \Lambda_2, \Lambda_3, \ldots$$

is said to converge, and to have as its limit the lattice Λ, *if there exist reduced bases*

$$Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_n^{(r)} \textit{ of } \Lambda_r \quad \textit{for} \quad r = 1, 2, 3, \ldots \tag{3·3}$$

and a basis $Y_1, Y_2, \ldots, Y_n$ *of* Λ,

such that $$\lim_{r \to \infty} |Y_g^{(r)} - Y_g| = 0, \quad \textit{where} \quad g = 1, 2, \ldots, n. \tag{3·4}$$

This definition implies that the points of Λ_r in any finite region independent of r tend to the points of Λ, as r tends to infinity.

From these two definitions is derived the following theorem which is fundamental for the study of star bodies:

THEOREM 2. *Every bounded infinite sequence of lattices contains a convergent infinite subsequence.*

Proof. Let $\Lambda_1, \Lambda_2, \Lambda_3, \ldots$ be any bounded sequence, and let $Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_n^{(r)}$ be a reduced basis of Λ_r for $r = 1, 2, 3, \ldots$, then from definition 1,

$$d(\Lambda_r) \leqslant c_1, \quad |Y_g^{(r)}| \geqslant c_2, \quad \text{where} \quad g = 1, 2, \ldots, n \text{ and } r = 1, 2, 3, \ldots, \tag{3·5}$$

and from theorem 1,

$$|Y_1^{(r)}|\,|Y_2^{(r)}| \ldots |Y_n^{(r)}| \leqslant \gamma_n d(\Lambda_r), \quad \text{where} \quad r = 1, 2, 3, \ldots, \tag{3·6}$$

hence
$$|Y_g^{(r)}| \leqslant \gamma_n c_1 c_2^{-(n-1)}, \quad \text{where} \quad g = 1, 2, \ldots, n \text{ and } r = 1, 2, 3, \ldots. \tag{3·7}$$

All co-ordinates of the basis points $Y_g^{(r)}$ $(g = 1, 2, \ldots, n;\ r = 1, 2, 3, \ldots)$ are therefore bounded, and so there exists an infinite sequence of indices

$$r_1, r_2, r_3, \ldots,$$

and a set of n points
$$Y_1, Y_2, \ldots, Y_n,$$

such that
$$\lim_{k\to\infty} |Y_g^{(r_k)} - Y_g| = 0, \quad \text{where} \quad g = 1, 2, \ldots, n, \tag{3·8}$$

whence
$$\lim_{k\to\infty} d(\Lambda_{r_k}) = \lim_{k\to\infty} |\{Y_1^{(r_k)}, Y_2^{(r_k)}, \ldots, Y_n^{(r_k)}\}| = |\{Y_1, Y_2, \ldots, Y_n\}|. \tag{3·9}$$

Further, from
$$\gamma_n d(\Lambda_{r_k}) \geqslant |Y_1^{(r_k)}|\,|Y_2^{(r_k)}| \ldots |Y_n^{(r_k)}| \geqslant c_2^n, \tag{3·10}$$

and
$$d(\Lambda_{r_k}) \geqslant \gamma_n^{-1} c_2^n, \tag{3·11}$$

it is deduced that
$$|\{Y_1, Y_2, \ldots, Y_n\}| \geqslant \gamma_n^{-1} c_2^n > 0, \tag{3·12}$$

and so the lattice Λ of basis $Y_1, Y_2, \ldots, Y_n$ satisfies the assertion.

4. DISTANCE FUNCTIONS AND STAR BODIES

DEFINITION 3. *A function*

$$F(X) = F(x_1, x_2, \ldots, x_n) \tag{4·1}$$

of the point $X = (x_1, x_2, \ldots, x_n)$ *in* R_n *is called a distance function if it satisfies the following conditions:*

(*a*) $F(X) \geqslant 0$ *for all points, and* $F(X) > 0$ *for at least one point;*

(*b*) $F(tX) = |t|\,F(X)$ *for all points* X *and all real numbers* t*; hence*

$$F(-X) = F(X) \quad \textit{and} \quad F(O) = 0;$$

(*c*) $F(X)$ *is a continuous function of* X.

DEFINITION 4. *The set K of all points X satisfying $F(X) \leqslant 1$ is called the star body of distance function $F(X)$; the subset C of all points of K with $F(X) = 1$ is called the boundary of K.*

It is evident that a star body K has the following properties:

(A) If X belongs to K, then tX, where $-1 \leqslant t \leqslant 1$, also belongs to K.

(B) The limit point of a convergent sequence of points of K also belongs to K.

(C) The origin O is an inner point of K; i.e. there exists a positive number ρ such that all points of the sphere $|X| \leqslant \rho$ belong to K.

For since $F(X)$ is continuous, it assumes on the sphere $|X| = 1$ a maximum value, say $1/\rho$. Then $F(X)|X|^{-1} \leqslant 1/\rho$ for all $X \neq O$, whence $F(X) \leqslant 1$, if X is a point of the sphere $|X| \leqslant \rho$.

THEOREM 3. *The star body K is bounded if and only if*

$$F(X) > 0 \textit{ for all points } X \neq O.$$

Proof. As a continuous function, $F(X)$ assumes on the sphere $|X| = 1$ a minimum, say μ. If $\mu = 0$, then $F(X)$ vanishes at a point $X \neq O$, and so it vanishes at all points of the line through O and X; hence K is not bounded. If, however, $\mu = 1/P > 0$, then $F(X)|X|^{-1} \geqslant 1/P$ for all $X \neq O$, hence $|X| \leqslant P$ if $F(X) \leqslant 1$, and so K is bounded.

5. THE TWO TYPES OF STAR BODIES

DEFINITION 5. *The lattice Λ is called K-admissible if $\Lambda - [O]$ contains no inner points of K.*

DEFINITION 6. *The star body K is called of the finite type if there exists at least one K-admissible lattice; it is called of the infinite type if no such lattice exists.*

THEOREM 4. *Every bounded star body is of the finite type.*

Proof. Let $P > 0$ be a number such that $|X| \leqslant P$ for all points of K, and denote by Λ the lattice of basis

$$X_1 = (P, 0, \ldots, 0),\ X_2 = (0, P, \ldots, 0),\ \ldots,\ X_n = (0, 0, \ldots, P). \tag{5·1}$$

Then $|X| \geqslant P$ for all points $X \neq O$ of Λ; hence Λ is K-admissible.

THEOREM 5. *Unbounded star bodies exist of the finite type, and also of the infinite type.*

Proof. (1) The star body K of distance function

$$F(X) = |x_1 x_2 \ldots x_n|^{1/n} \tag{5·2}$$

is not bounded. To show that K is of the finite type, denote by $\mathfrak{K}$ any totally real algebraic field of degree n, by

$$\omega_1^{(g)}, \omega_2^{(g)}, \ldots, \omega_n^{(g)}, \quad \text{where} \quad g = 1, 2, \ldots, n,$$

conjugate integral bases of the n fields $\mathfrak{K}^{(1)}, \mathfrak{K}^{(2)}, \ldots, \mathfrak{K}^{(n)}$ conjugate to $\mathfrak{K}$, and by Λ the lattice of basis

$$X_h = (\omega_h^{(1)}, \omega_h^{(2)}, \ldots, \omega_h^{(n)}), \quad \text{where} \quad h = 1, 2, \ldots, n. \tag{5·3}$$

Then, except for the sign, $F(X)$ is the norm of an integer $\alpha \neq 0$ in $\mathfrak{K}$ if X lies in $\Lambda - [O]$; hence $F(X) \geqslant 1$ for all lattice points $X \neq O$.

(2) The star body of distance function

$$F(X) = |x_1^2 x_2 \ldots x_n|^{1/(n+1)} \tag{5·4}$$

likewise is not bounded, but it is of the infinite type. For let Λ be any lattice, and denote by $t_1, t_2, \ldots, t_n$, n positive numbers of product $d(\Lambda)$. By Minkowski's theorem on linear forms, there exists a point $X = (x_1, x_2, \ldots, x_n) \neq O$ of Λ such that

$$|x_1| \leqslant t_1, \quad |x_2| \leqslant t_2, \quad \ldots, \quad |x_n| \leqslant t_n, \tag{5·5}$$

hence

$$0 < F(X) \leqslant \{t_1 d(\Lambda)\}^{1/(n+1)}. \tag{5·6}$$

If it be assumed now that $t_1 < d(\Lambda)^{-1}$, then X is an inner point of K. Therefore Λ is not K-admissible.

Unless otherwise stated, all star bodies considered are from now on assumed to be of the finite type.

6. THE DETERMINANT OF A STAR BODY

Let K: $F(X) \leqslant 1$, be a star body of the finite type. By definition 6, the set $\Lambda(K)$ of all K-admissible lattices is not empty. Hence the lower bound

$$\Delta(K) = \text{l.b.}\, d(\Lambda) \tag{6·1}$$

extended over all elements of $\Lambda(K)$, exists; $\Delta(K)$ is called the determinant of K. For star bodies K of the infinite type, put $\Delta(K) = \infty$.

THEOREM 6. *The determinant of a star body is positive.*

Proof. By the property (C) of a star body (§ 4), K contains the sphere $|X| \leqslant \rho$, hence also the cube

$$\max(|x_1|, |x_2|, \ldots, |x_n|) \leqslant \rho n^{-\frac{1}{2}}. \tag{6·2}$$

By Minkowski's theorem on linear forms, every lattice of determinant

$$d(\Lambda) < \rho^n n^{-\frac{1}{2}n}$$

contains an inner point $X \neq O$ of this cube, i.e. of K, and so such a lattice cannot be K-admissible. Hence, for every K-admissible lattice Λ,

$$d(\Lambda) \geqslant \rho^n n^{-\frac{1}{2}n}, \tag{6·3}$$

whence

$$\Delta(K) \geqslant \rho^n n^{-\frac{1}{2}n} > 0. \tag{6·4}$$

THEOREM 7. *If the star body H is contained in the star body K, then*

$$\Delta(H) \leqslant \Delta(K). \tag{6·5}$$

Proof. Every K-admissible lattice is also H-admissible.

7. THE EXISTENCE OF A CRITICAL LATTICE

DEFINITION 7. *The lattice Λ is called a critical lattice of K if it is K-admissible and $d(\Lambda) = \Delta(K)$.*

The following theorem is fundamental for the theory:

THEOREM 8. *Every star body of the finite type possesses at least one critical lattice.*

Proof. From the definition of $\Delta(K)$, there exists an infinite sequence of K-admissible lattices

$$\Lambda_1, \Lambda_2, \Lambda_3, \ldots,$$

not necessarily all different, such that

$$\lim_{r\to\infty} d(\Lambda_r) = \Delta(K); \tag{7·1}$$

it may be assumed further, without loss of generality, that

$$d(\Lambda_r) \leqslant 2\Delta(K), \quad \text{where} \quad r = 1, 2, 3, \ldots. \tag{7·2}$$

Moreover, since the sphere $|X| \leqslant \rho$ is contained in K,

$$|X| \geqslant \rho \quad \text{for all points } X \neq O \text{ of } \Lambda_r, \quad \text{where} \quad r = 1, 2, 3, \ldots. \tag{7·3}$$

From (7·2) and (7·3) the sequence $\{\Lambda_r\}$ is bounded, and hence, from theorem 2, it contains a convergent infinite subsequence

$$\Lambda_{r_1}, \Lambda_{r_2}, \Lambda_{r_3}, \ldots,$$

say of limit Λ. Denote by $Y_1^{(r_k)}, Y_2^{(r_k)}, \ldots, Y_n^{(r_k)}$ a reduced basis of Λ_{r_k}, by $Y_1, Y_2, \ldots, Y_n$ a basis of Λ, taken such that

$$\lim_{k\to\infty} | Y_g^{(r_k)} - Y_g | = 0, \quad \text{where} \quad g = 1, 2, \ldots, n, \tag{7·4}$$

hence

$$d(\Lambda) = |\{Y_1, Y_2, \ldots, Y_n\}| = \lim_{k\to\infty} |\{Y_1^{(r_k)}, Y_2^{(r_k)}, \ldots, Y_n^{(r_k)}\}| = \lim_{k\to\infty} d(\Lambda_{r_k}) = \Delta(K). \tag{7·5}$$

Let further

$$Y = u_1 Y_1 + \ldots + u_n Y_n \neq O, \quad \text{where} \quad u_1, \ldots, u_n \text{ are integers} \tag{7·6}$$

be any point of Λ, and put

$$Y^{(r_k)} = u_1 Y_1^{(r_k)} + \ldots + u_n Y_n^{(r_k)}, \quad \text{where} \quad k = 1, 2, 3, \ldots; \tag{7·7}$$

then

$$\lim_{k\to\infty} | Y^{(r_k)} - Y | = 0. \tag{7·8}$$

Hence $Y^{(r_k)} \neq O$ for sufficiently large k, and so $F(Y^{(r_k)}) \geqslant 1$ since Λ_{r_k} is K-admissible, whence

$$F(Y) = \lim_{k\to\infty} F(Y^{(r_k)}) \geqslant 1. \tag{7·9}$$

From (7·5) and (7·9), Λ satisfies the assertion.

8. THE CONTINUITY OF $\Delta(K)$

If $K: F(X) \leqslant 1$, is any star body, and if t is a positive number, then we denote by tK the star body of distance function $t^{-1}F(X)$, i.e. the set of all points X for which $F(X) \leqslant t$. From homogeneity, it is evident that

$$\Delta(tK) = t^n \Delta(K). \tag{8·1}$$

The set of all points X in K for which $|X| \leqslant t$ is further denoted by K^t.

THEOREM 9. *Let $K, K_1, K_2, \ldots$ be an infinity of star bodies of the finite type, satisfying the following conditions:*

(a) To every $\epsilon > 0$, there is a positive integer $N(\epsilon)$ such that K_r is contained in $(1+\epsilon)K$ if $r \geqslant N(\epsilon)$.

(b) To every $t > 0$ and every $\epsilon > 0$, there is a positive integer $N(t, \epsilon)$ such that K^t is contained in $(1+\epsilon)K_r$ if $r \geqslant N(t, \epsilon)$.

Then
$$\lim_{r\to\infty} \Delta(K_r) = \Delta(K). \tag{8·2}$$

Proof. From (*a*), by theorem 7,

$$\Delta(K_r) \leqslant \Delta((1+\epsilon)K) = (1+\epsilon)^n \Delta(K), \tag{8·3}$$

whence for $\epsilon \to 0$,
$$\limsup_{r\to\infty} \Delta(K_r) \leqslant \Delta(K). \tag{8·4}$$

It will now be shown that also

$$\liminf_{r\to\infty} \Delta(K_r) \geqslant \Delta(K). \tag{8·5}$$

Let this inequality be false. Then there exists an infinite sequence of indices $r_1, r_2, r_3, \ldots$ not smaller than $N(\rho, 1)$ such that

$$\Delta(K_{r_k}) \leqslant 2\Delta(K), \quad \text{and} \quad \lim_{k\to\infty} \Delta(K_{r_k}) < \Delta(K). \tag{8·6}$$

Denote by Λ_{r_k} a critical lattice of K_{r_k}; therefore

$$d(\Lambda_{r_k}) \leqslant 2\Delta(K). \tag{8·7}$$

Then from (*b*) above, on taking $t = \rho$, $\epsilon = 1$, the star body $2K_{r_k}$ contains K^ρ, i.e. the sphere $|X| \leqslant \rho$; hence K_{r_k} contains the sphere $|X| \leqslant \frac{1}{2}\rho$. Since Λ_{r_k} is K_{r_k}-admissible, this implies that

$$|X| \geqslant \tfrac{1}{2}\rho \text{ for all points } X \neq O \text{ of } \Lambda_{r_k}.$$

It is clear from this and (8·7) that the sequence of lattices $\{\Lambda_{r_k}\}$ is bounded. Therefore, from theorem 2, this sequence contains a convergent infinite subsequence

$$\Lambda^{(1)} = \Lambda_{r_{k_1}}, \Lambda^{(2)} = \Lambda_{r_{k_2}}, \Lambda^{(3)} = \Lambda_{r_{k_3}}, \ldots, \tag{8·8}$$

of limit Λ, say. For shortness write

$$K^{(1)} = K_{r_{k_1}}, K^{(2)} = K_{r_{k_2}}, K^{(3)} = K_{r_{k_3}}, \ldots, \tag{8·9}$$

then, as in the proof of the last theorem,

$$d(\Lambda) = \lim_{l\to\infty} d(\Lambda^{(l)}) = \lim_{l\to\infty} \Delta(K^{(l)}), \tag{8·10}$$

and so

$$d(\Lambda) = \lim_{k\to\infty} \Delta(K_{r_k}) < \Delta(K). \tag{8·11}$$

This means that Λ is not K-admissible; hence Λ contains at least one point $Y \neq O$ which is an inner point of K.

Denote now by $Y_1^{(l)}, Y_2^{(l)}, \ldots, Y_n^{(l)}$ a reduced basis of $\Lambda^{(l)}$, and by $Y_1, Y_2, \ldots, Y_n$ a basis of Λ taken such that

$$\lim_{l\to\infty} | Y_g^{(l)} - Y_g | = 0, \quad \text{where} \quad g = 1, 2, \ldots, n; \tag{8·12}$$

then Y can be written as

$$Y = u_1 Y_1 + \ldots + u_n Y_n \tag{8·13}$$

with integral coefficients $u_1, \ldots, u_n$ not all zero. Now put

$$Y^{(l)} = u_1 Y_1^{(l)} + \ldots + u_n Y_n^{(l)}, \tag{8·14}$$

then $Y^{(l)}$ belongs to $\Lambda^{(l)}$,

$$Y^{(l)} \neq O, \quad \text{and} \quad \lim_{l\to\infty} | Y^{(l)} - Y | = 0, \tag{8·15}$$

whence, for sufficiently large indices l,

$$| Y^{(l)} | \leqslant 2 | Y |. \tag{8·16}$$

Since Y is an inner point of K and different from O, there is an $\epsilon > 0$ such that

$$F(Y) \leqslant \frac{1}{1+3\epsilon}, \tag{8·17}$$

hence, if l is sufficiently large, from (8·15) it follows that

$$F(Y^{(l)}) \leqslant \frac{1}{1+2\epsilon}, \tag{8·18}$$

and so $(1+2\epsilon)\, Y^{(l)} \neq O$ belongs to K. This implies, from (8·16), that $(1+2\epsilon)\, Y^{(l)}$ is a point of K^t, where $t = 2(1+2\epsilon) | Y |$. Hence, from (*b*) above, the point $(1+2\epsilon)\, Y^{(l)}$ belongs to $(1+\epsilon) K^{(l)}$ if l is sufficiently large. This implies that $Y^{(l)}$ is a point of $\dfrac{1+\epsilon}{1+2\epsilon} K^{(l)}$ and so is an inner point of $K^{(l)}$. However, this is impossible since $Y^{(l)} \neq O$ and since $\Lambda^{(l)}$ is a critical lattice of $K^{(l)}$.

THEOREM 10. *Let $K: F(X) \leqslant 1$ be a star body of the finite type, $G(X)$ an arbitrary distance function, and t a positive parameter. Then the star body*

$$K_t: F_t(X) \leqslant 1, \quad \textit{where} \quad F_t(X) = \max(F(X), t^{-1}G(X)),$$

is also of the finite type, and further

$$\lim_{t\to\infty} \Delta(K_t) = \Delta(K). \tag{8·19}$$

Proof. It is evident from definition 3 that $F_t(X)$ is a distance function. Since $F_t(X) \geqslant F(X)$ for all X and t, K_t is contained in K and so is a star body of the finite type. Further, since the set H: $G(X) \leqslant 1$ is a star body, there exists a number $\tau > 0$ such that H contains the whole sphere $|X| \leqslant \tau$. The sphere $|X| \leqslant \tau t$ is then contained in tH, and so $K^{\tau t}$, which is a subset of this sphere, is contained in K_t. The hypothesis of theorem 9 is therefore satisfied, and so

$$\Delta(K) = \lim_{r\to\infty} \Delta(K_{t_r}) \tag{8·20}$$

for every sequence of positive numbers $t_1, t_2, t_3, \ldots$ of limit infinity. This proves the assertion.

The last theorem, for $G(X) = |X|$, shows that

$$\Delta(K) = \lim_{t\to\infty} \Delta(K^t). \tag{8·21}$$

Originally (Mahler 1943), I used this formula as the definition of $\Delta(K)$ for unbounded star bodies, so reducing the problem to one for the bounded case.

Remark. The results of this paragraph remain true when $\Delta(K) = \infty$.

9. LATTICE POINTS ON THE BOUNDARY OF A BOUNDED STAR BODY

THEOREM 11. *If K is a bounded star body, then every critical lattice of K has n independent points on the boundary C of K.*

Proof. Let Λ be a K-admissible lattice which does not contain n independent points on C. Then denote by Π the set of all points of Λ on C, and by L the linear space of lowest dimension f $(0 \leqslant f \leqslant n-1)$ containing Π. By Minkowski's method of adaptation of lattices, a basis $Y_1, \ldots, Y_n$ of Λ can be found such that $Y_1, \ldots, Y_f$ lie in and generate L, while $Y_{f+1}, \ldots, Y_n$ lie outside L. Let $\epsilon > 0$ be sufficiently small and denote by Λ^* the lattice of basis $Y_1, \ldots, Y_f$, $(1-\epsilon)Y_{f+1}, \ldots, (1-\epsilon)Y_n$. This lattice is K-admissible since O and the elements of Π are its only points belonging to K. Since $d(\Lambda^*) = (1-\epsilon)^{n-f} d(\Lambda) < d(\Lambda)$, Λ^* is of smaller determinant than Λ, and so Λ is not critical.

This theorem shows that in the case of a bounded star body K, every critical lattice Λ has at least $2n$ points on its boundary C, namely, n independent points $P_1, \ldots, P_n$ and their images $-P_1, \ldots, -P_n$ in O. If

$$\pm P_1, \pm P_2, \ldots, \pm P_n$$

are the only points on C of the lattice Λ, then Λ is called a singular lattice of K; otherwise it is called a regular lattice. The example in the next paragraph shows that star bodies with singular lattices do exist.

10. AN EXAMPLE OF A STAR BODY WITH A SINGULAR LATTICE

THEOREM 12. *There exists a bounded star body with just one critical lattice. Moreover, this lattice is singular.*

Proof. Let ϵ be so small a positive constant that

$$(1-\epsilon)^n > \tfrac{99}{100}, \quad (1-\epsilon)^{n-1}\sqrt[n]{\tfrac{3}{2}} > 1, \quad \epsilon < n^{-1}\left(\sqrt[n]{\left(\tfrac{11}{10}\right)} - 1\right), \tag{10·1}$$

and let $Q \neq O$ be a point in R_n. The set $S_\epsilon(Q)$ of all points

$$P = tQ + (t-1)\,\epsilon R, \quad \text{where} \quad t \geqslant 1 \text{ and } |R| \leqslant 1, \tag{10·2}$$

is a cone with vertex at Q and its open side away from O. For when t is fixed and R describes all points of the unit sphere $|R| \leqslant 1$, then P lies on or in a sphere centre at tQ and radius $(t-1)\,\epsilon$; on varying t, we obtain $S_\epsilon(Q)$ as the sum set of these spheres.

Denote further by Λ_0 the lattice of all points with integral co-ordinates, i.e. of basis

$$P_1 = (1, 0, \ldots, 0),\ P_2 = (0, 1, \ldots, 0),\ \ldots,\ P_n = (0, 0, \ldots, 1), \tag{10·3}$$

and of determinant $d(\Lambda_0) = 1$.

The cube $$W: \quad |x_1| \leqslant \sqrt[n]{\tfrac{3}{2}},\ |x_2| \leqslant \sqrt[n]{\tfrac{3}{2}},\ \ldots,\ |x_n| \leqslant \sqrt[n]{\tfrac{3}{2}} \tag{10·4}$$

contains 3^n points of Λ_0, namely, the origin O, the $2n$ points $\pm P_1, \pm P_2, \ldots, \pm P_n$, and the m points $P'_1, P'_2, \ldots, P'_m$, where

$$P'_h = (x_1^{(h)}, x_2^{(h)}, \ldots, x_n^{(h)}), \qquad x_g^{(h)} = 0, 1, \text{or } -1, \quad \sum_{g=1}^{n} |x_g^{(h)}| \geqslant 2. \tag{10·5}$$

Denote by K the set of all those points of W which are not inner points of one of the cones

$$S_\epsilon(\pm P_g), \quad \text{where} \quad g = 1, 2, \ldots, n, \qquad \text{or} \qquad S_\epsilon[(1-\epsilon)\,P'_h], \quad \text{where} \quad h = 1, 2, \ldots, m.$$

Then K is a bounded star body, and the cube

$$V: \quad |x_1| \leqslant 1-\epsilon,\ |x_2| \leqslant 1-\epsilon,\ \ldots,\ |x_n| \leqslant 1-\epsilon, \tag{10·6}$$

obviously is a subset of K. Therefore from theorem 7, Minkowski's theorem on linear forms, and from (10·1)

$$\Delta(K) \geqslant \Delta(V) = (1-\epsilon)^n > \tfrac{5}{6}. \tag{10·7}$$

On the other hand $$\Delta(K) \leqslant d(\Lambda_0) = 1, \tag{10·8}$$

since, by the construction, Λ_0 is K-admissible. Hence, if Λ is any critical lattice of K, then

$$\tfrac{5}{6} < d(\Lambda) \leqslant 1. \tag{10·9}$$

Each one of the n parallelepipeds

$$U_g: \quad |x_g| \leqslant \sqrt[n]{\tfrac{3}{2}}, \quad |x_l| \leqslant 1-\epsilon \quad \text{for} \quad l = 1, 2, \ldots, g-1, g+1, \ldots, n \tag{10·10}$$

from (10·1) is of volume $$2^n(1-\epsilon)^{n-1}\sqrt[n]{\tfrac{3}{2}} > 2^n. \tag{10·11}$$

Hence, from Minkowski's theorem on linear forms, at least one point of $\Lambda - [O]$ is an inner point of U_g, say the point $P_g^* = (\xi_1^{(g)}, \xi_2^{(g)}, \ldots, \xi_n^{(g)})$. This point lies in one of the two cones $S_\epsilon(\pm P_g)$, since the other inner points of U_g are also inner points of K. There is no loss of generality in assuming that P_g^* belongs to $S_\epsilon(P_g)$ and so may be written as

$$P_g^* = t_g P_g + (t_g - 1)\,\epsilon R_g, \quad \text{where} \quad t_g \geqslant 1 \text{ and } |R_g| \leqslant 1. \tag{10·12}$$

Therefore if, say, $R_g = (\eta_1^{(g)}, \eta_2^{(g)}, \dots, \eta_n^{(g)})$, then

$$\xi_g^{(g)} = t_g + (t_g - 1)\,\epsilon\eta_g^{(g)} \tag{10·13}$$

and

$$\eta_g^{(g)} \geqslant -1; \tag{10·14}$$

and since P_g lies in U_g,

$$\sqrt[n]{\tfrac{3}{2}} \geqslant \xi_g^{(g)} = t_g + (t_g - 1)\,\epsilon\eta_g^{(g)} \geqslant t_g - (t_g - 1)\,\epsilon > (1-\epsilon)\,t_g, \tag{10·15}$$

whence

$$1 \leqslant t_g < \frac{\sqrt[n]{\frac{3}{2}}}{1-\epsilon}, \quad \text{where} \quad g = 1, 2, \dots, n. \tag{10·16}$$

Denote now by D the determinant

$$D = \{P_1^*, P_2^*, \dots, P_n^*\}, \tag{10·17}$$

by E the unit determinant

$$E = \{P_1, P_2, \dots, P_n\} = d(\Lambda_0) = +1, \tag{10·18}$$

and by $E(g_1, g_2, \dots, g_r)$, where $1 \leqslant r \leqslant n$, $1 \leqslant g_1 < g_2 < \dots < g_r < n$,

the determinant which is obtained from E if the points $P_{g_1}, P_{g_2}, \dots, P_{g_r}$ in it are replaced by the points $R_{g_1}, R_{g_2}, \dots, R_{g_r}$ of the same indices. Obviously $E(g_1, g_2, \dots, g_r)$ is equal to its cofactor of order r belonging to the rows and columns of indices $g_1, g_2, \dots, g_r$. Hence

$$|\,E(g_1, g_2, \dots, g_r)\,| \leqslant r! \tag{10·19}$$

since the moduli of the co-ordinates of $R_{g_1}, R_{g_2}, \dots, R_{g_r}$ are not larger than 1, and since a determinant of order r consists of $r!$ terms.

From (10·12), D can be split into a sum of 2^n determinants, namely,

$$D = t_1 t_2 \dots t_n \left(E + \sum_{r=1}^{n}{}^{*} E(g_1, g_2, \dots, g_r)\, \epsilon^r \frac{t_{g_1}-1}{t_{g_1}} \frac{t_{g_2}-1}{t_{g_2}} \dots \frac{t_{g_r}-1}{t_{g_r}} \right), \tag{10·20}$$

with the abbreviation

$$\sum_{r=1}^{n}{}^{*} = \sum_{r=1}^{n} \sum_{\substack{g_1, g_2, \dots, g_r = 1 \\ g_1 < g_2 < \dots < g_r}}^{n}. \tag{10·21}$$

Now from (10·1) and (10·19)

$$|\,E(g_1, g_2, \dots, g_r)\,|\, \epsilon^r \leqslant r!\, \epsilon^r \leqslant (r\epsilon)^r \leqslant (n\epsilon)^r \leqslant \{\sqrt[n]{(\tfrac{11}{10})} - 1\}^r, \tag{10·22}$$

hence

$$\left| \sum_{r=1}^{n}{}^{*} E(g_1, g_2, \dots, g_r)\, \epsilon^r \frac{t_{g_1}-1}{t_{g_1}} \dots \frac{t_{g_r}-1}{t_{g_r}} \right| \leqslant \sum_{r=1}^{n}{}^{*} \{\sqrt[n]{(\tfrac{11}{10})} - 1\}^r \frac{t_{g_1}-1}{t_{g_1}} \dots \frac{t_{g_r}-1}{t_{g_r}}$$

$$= \prod_{g=1}^{n} \left(1 + \{\sqrt[n]{(\tfrac{11}{10})} - 1\} \frac{t_g - 1}{t_g} \right) - 1 = (t_1 t_2 \dots t_n)^{-1} \prod_{g=1}^{n} \{1 + \sqrt[n]{(\tfrac{11}{10})}\,(t_g - 1)\} - 1, \tag{10·23}$$

whence

$$D \leqslant \prod_{g=1}^{n} \{1 + \sqrt[n]{(\tfrac{11}{10})}\,(t_g - 1)\}, \tag{10·24}$$

and

$$D \geqslant 2 \prod_{g=1}^{n} t_g - \prod_{g=1}^{n} \{1 + \sqrt[n]{(\tfrac{11}{10})}\,(t_g - 1)\}. \tag{10·25}$$

From (10·1), (10·16) and (10·24) then

$$D < \prod_{g=1}^{n} \{\sqrt[n]{(\tfrac{11}{10})} + \sqrt[n]{(\tfrac{11}{10})}\,(t_g - 1)\} = \prod_{g=1}^{n} \{\sqrt[n]{(\tfrac{11}{10})}\, t_g\} < \prod_{g=1}^{n} \left(\frac{\sqrt[n]{(\tfrac{11}{10})}\,\sqrt[n]{(\tfrac{3}{2})}}{1-\epsilon}\right) = \tfrac{33}{20}(1-\epsilon)^{-n} < \tfrac{5}{3}. \tag{10·26}$$

Further, since $2 - (\frac{11}{10})^{r/n} > 0$ for $r = 1, 2, \ldots, n$, then from (10·16) and (10·25),

$$D \geqslant 2 \prod_{g=1}^{n} \{1 + (t_g - 1)\} - \prod_{g=1}^{n} \{1 + \sqrt[n]{(\tfrac{11}{10})}\,(t_g - 1)\}$$

$$= 1 + \sum_{r=1}^{n}{}^{*} \{2 - (\tfrac{11}{10})^{r/n}\}\,(t_{g_1} - 1) \ldots (t_{g_r} - 1) \geqslant 1, \tag{10·27}$$

with $D = 1$ if and only if $t_1 = t_2 = \ldots = t_n = 1$.

This proves that

$$1 \leqslant D < \tfrac{5}{3}, \tag{10·28}$$

the lower bound being assumed if and only if $t_1 = t_2 = \ldots = t_n = 1$, i.e. if

$$P_1^* = P_1,\ P_2^* = P_2,\ \ldots,\ P_n^* = P_n.$$

Since $D > 0$, the n points $P_1^*, P_2^*, \ldots, P_n^*$ are independent; therefore

$$D = j d(\Lambda), \tag{10·29}$$

where j is a positive integer. From (10·9) and (10·28) it follows that

$$\tfrac{5}{3} > j \cdot \tfrac{5}{6}, \quad j < 2, \tag{10·30}$$

and so $j = 1$, $d(\Lambda) = D \geqslant 1$, with equality if and only if $\Lambda = \Lambda_0$. Since Λ_0 is K-admissible and since $d(\Lambda_0) = 1$, this completes the proof that Λ_0 is the only critical lattice of K, and also that Λ_0 is singular.

COROLLARY. *For any given integer $m \geqslant n$, there exists a bounded star body K with a critical lattice having just $2m$ points on the boundary of K.*

Proof. Nearly obvious, because any star body K' has the required property if it satisfies the following three conditions: (*a*) K, as defined in the last proof, is a subset of K'. (*b*) Λ_0, as defined in the last proof, is K'-admissible. (*c*) Just $2m$ points of Λ_0 lie on the boundary of K'.

Remark. In an earlier paper on star domains,* I discussed a method by which to obtain $\Delta(K)$ and the critical lattices for every bounded two-dimensional star body, provided the boundary consists of a finite number of analytical arcs. This method may be extended to the n-dimensional case, but, naturally, the calculations now become very complicated.

11. THE LATTICE FUNCTION $F(\Lambda)$

If Λ is a lattice, t a positive number, and $t\Lambda$ denotes the lattice of all points tP where P runs over Λ, it is obvious that

$$d(t\Lambda) = t^n d(\Lambda). \tag{11·1}$$

---

* Mahler—On lattice points in two-dimensional star domains, to appear in the *Proceedings of the London Mathematical Society.*

Further, if K denotes the star body (not necessarily bounded) of distance function $F(X)$, write

$$F(\Lambda) = \text{l.b.}\, F(P), \tag{11·2}$$

for the lower bound of $F(P)$ extended over all points $P \neq O$ of Λ. Then the symbol $F(\Lambda)$ has the following evident properties:

Λ is K-admissible if and only if $F(\Lambda) \geqslant 1$.

Λ is a critical lattice of K if and only if $F(\Lambda) = 1$, $d(\Lambda) = \Delta(K)$;

further

$$F(t\Lambda) = tF(\Lambda) \quad \text{if} \quad t > 0. \tag{11·3}$$

A star body is therefore of the finite type if $F(\Lambda) > 0$ for at least one lattice, and is of the infinite type if $F(\Lambda) = 0$ for all lattices.

In the special case when K is a bounded star body, it is easily seen that $F(\Lambda)$ is a continuous function of Λ; i.e. if $\Lambda_1, \Lambda_2, \Lambda_3, \ldots$ is a convergent sequence of lattices of limit Λ, then

$$\lim_{r\to\infty} F(\Lambda_r) = F(\Lambda). \tag{11·4}$$

If, however, K is an unbounded star body, then $F(\Lambda)$ need not be continuous, as the following example shows. We choose

$$F(X) = |x_1 x_2 \ldots x_n|^{1/n}, \tag{11·5}$$

and take for Λ the lattice of basis

$$X_h = (\omega_h^{(1)}, \omega_h^{(2)}, \ldots, \omega_h^{(n)}), \quad \text{where} \quad h = 1, 2, \ldots, n, \tag{11·6}$$

as defined in the proof of part (1) of theorem 5; there is no restriction in assuming that this basis is reduced. Further, denote by

$$X_1^{(r)}, X_2^{(r)}, \ldots, X_n^{(r)}, \quad \text{where} \quad r = 1, 2, 3, \ldots,$$

an infinity of sets of n independent points with rational co-ordinates such that

$$\lim_{r\to\infty} |X_h^{(r)} - X_h| = 0, \quad \text{where} \quad h = 1, 2, \ldots, n, \tag{11·7}$$

and such that further $X_1^{(r)}, X_2^{(r)}, \ldots, X_n^{(r)}$ form a reduced basis of the lattice Λ_r generated by these n points. Then by the proof of theorem 5,

$$F(\Lambda) \geqslant 1, \tag{11·8}$$

while, on the other hand,

$$F(\Lambda_r) = 0 \tag{11·9}$$

and

$$\lim_{r\to\infty} F(\Lambda_r) = 0, \tag{11·10}$$

since a linear form with rational coefficients represents zero.

12. LATTICE POINTS NEAR THE BOUNDARY OF AN UNBOUNDED STAR BODY

It was seen in § 9 that a critical lattice of any bounded star body has at least $2n$ points on its boundary. For unbounded star bodies, this is no longer so; as will be seen in the next paragraph, there exists an unbounded star body of the finite type such that at least one of its critical lattices has no point on its boundary.

It may then be asked, however, whether lattice points lie arbitrarily near to the boundary of K. The answer is given by the nearly obvious

THEOREM 13. *If* K: $F(X) \leqslant 1$ *is a star body of the finite type and* Λ *is a critical lattice of* K, *then to every* $\epsilon > 0$ *there exists a point* P *of* Λ *such that*

$$1 \leqslant F(P) < 1+\epsilon. \tag{12·1}$$

Proof. If $F(P) \geqslant 1+\epsilon$ for every point $P \neq O$ of Λ, then

$$F(\Lambda) \geqslant 1+\epsilon, \tag{12·2}$$

whence

$$F\left(\frac{\Lambda}{1+\epsilon}\right) \geqslant 1. \tag{12·3}$$

Therefore $\frac{\Lambda}{1+\epsilon}$ is also K-admissible, but is of smaller determinant than Λ, and so Λ is not critical.

This theorem leads to:

PROBLEM A. *Let* K: $F(X) \leqslant 1$ *be a star body of the finite type,* Λ *a critical lattice of* K, *and* $\epsilon > 0$ *any arbitrarily small number. Do there exist n independent points* $P_1, P_2, \ldots, P_n$ *of* Λ *such that*

$$1 \leqslant F(P_g) < 1+\epsilon, \quad \textit{where} \quad g = 1, 2, \ldots, n? \tag{12·4}$$

I have not been able to decide this question. The difficulty lies in the fact that $F(\Lambda)$ may be discontinuous, and so the method of the proof of theorem 11 cannot be applied.

Remark. From theorems 8 and 13, for any given $\epsilon > 0$, every lattice of determinant $d(\Lambda) = \Delta(K)$ contains a point $P \neq O$ satisfying $F(P) < 1+\epsilon$.

13. AN EXAMPLE OF AN UNBOUNDED STAR BODY WITH NO CRITICAL LATTICE POINTS ON ITS BOUNDARY

THEOREM 14. *Let* $F_0(X)$ *be the distance function*

$$F_0(X) = |x_1 x_2 \ldots x_n|^{1/n}, \tag{13·1}$$

and let further $F(X)$ *be any distance function satisfying the conditions*

$$F(X) \geqslant F_0(X) \quad \textit{if} \quad F_0(X) > 0, \tag{13·2}$$

$$\frac{F(X)}{F_0(X)} \to 1 \quad \textit{if} \quad F_0(X) > 0,\ |X|^{-1} F_0(X) \to 0. \tag{13·3}$$

Denote by K_0 *and* K *the star bodies of distance functions* $F_0(X)$ *and* $F(X)$, *respectively. Then*

$$\Delta(K) = \Delta(K_0). \tag{13·4}$$

Proof. K is a subset of K_0, and so from theorem 7, it follows that

$$\Delta(K) \leqslant \Delta(K_0). \tag{13·5}$$

Now assume that $$\Delta(K) < \Delta(K_0); \tag{13·6}$$

this assumption leads to a contradiction, as will be proved.

The function $f(X)$ defined by

$$f(X) = \left.\begin{cases} \dfrac{F(X)}{F_0(X)} & \text{if } F_0(X) \neq 0, \\ 1 & \text{if } F_0(X) = 0,\ X \neq O, \end{cases}\right\} \tag{13·7}$$

and not defined if $X = O$, is continuous and therefore bounded for all points of the unit sphere $|X| = 1$. Let $c \geqslant 1$ be its upper bound on this sphere:

$$f(X) \leqslant c \quad \text{if} \quad |X| = 1. \tag{13·8}$$

Then, since $$f(tX) = f(X) \quad \text{for} \quad t \neq 0, \tag{13·9}$$

c is the upper bound of $f(X)$ for all $X \neq O$, therefore

$$F(X) \leqslant cF_0(X) \tag{13·10}$$

for all X, since this inequality remains true if $X = O$.

Let now Λ be any critical lattice of K; then, from (13·6),

$$d(\Lambda) < \Delta(K_0), \tag{13·11}$$

or, say, $$d(\Lambda) = (1+\alpha)^{-(n+1)}\Delta(K_0), \tag{13·12}$$

where α is some positive number. Put

$$(1+\alpha)\Lambda = \Lambda', \tag{13·13}$$

so that Λ' is $(1+\alpha)$ K-admissible, and

$$d(\Lambda') = (1+\alpha)^{-1}\Delta(K_0) < \Delta(K_0). \tag{13·14}$$

Denote further by Σ the set of all points of Λ' which are inner points of K_0. If P is any point of Σ, then

$$F(P) \geqslant 1+\alpha, \quad F_0(P) < 1, \tag{13·15}$$

whence $$\frac{F(P)}{F_0(P)} > 1+\alpha, \tag{13·16}$$

and further, from (13·10),

$$F_0(P) \geqslant \frac{1}{c}F(P) \geqslant \frac{1+\alpha}{c} > 0. \tag{13·17}$$

But from (13·3) there exists a positive number β such that

$$\left|\frac{F(X)}{F_0(X)} - 1\right| \leqslant \alpha \quad \text{if} \quad F_0(X) \neq 0,\ |X|^{-1}F_0(X) < \beta. \tag{13·18}$$

Hence, by the inequalities just proved,

$$|P|^{-1}F_0(P) \geqslant \beta, \tag{13·19}$$

and so $$|P| \leqslant \frac{F_0(P)}{\beta} < \frac{1}{\beta}. \tag{13·20}$$

Next, if $P = (p_1, p_2, \ldots, p_n)$, then

$$|p_1 p_2 \ldots p_n| = F_0(P)^n \geqslant \left(\frac{1+\alpha}{c}\right)^n, \qquad (13\cdot21)$$

and
$$\max(|p_1|, |p_2|, \ldots, |p_n|) \leqslant |P| < \frac{1}{\beta}; \qquad (13\cdot22)$$

and so, finally,
$$|p_1| \geqslant |p_2 \ldots p_n|^{-1}\left(\frac{1+\alpha}{c}\right)^n > \left(\frac{1+\alpha}{c}\right)^n \beta^{n-1}. \qquad (13\cdot23)$$

Denote by r any positive integer, and by $Y = \Omega_r X$ the unimodular linear transformation

$$y_1 = r^{n-1}x_1, \quad y_2 = r^{-1}x_2, \ldots, y_n = r^{-1}x_n. \qquad (13\cdot24)$$

Further denote by $\Lambda_r = \Omega_r \Lambda'$ the lattice of all points $Q = \Omega_r P$ where P runs over Λ', and by $\Sigma_r = \Omega_r \Sigma$ the set of all points $Q = \Omega_r P$ where P lies in Σ. Then obviously

$$d(\Lambda_r) = d(\Lambda'), \qquad (13\cdot25)$$

and Σ_r consists of all and only all those points of Λ_r which are inner points of K_0.

If $P = (p_1, p_2, \ldots, p_n)$ is a point of Σ and $Q = \Omega_r P = (q_1, q_2, \ldots, q_n)$ is the corresponding point of Σ_r, then, from (13·23)

$$|q_1| = r^{n-1}|p_1| > \left(\frac{1+\alpha}{c}\right)^n (\beta r)^{n-1}, \qquad (13\cdot26)$$

and so
$$|Q| \geqslant |q_1| > \left(\frac{1+\alpha}{c}\right)^n (\beta r)^{n-1}. \qquad (13\cdot27)$$

As in § 8, denote by K_0^t, where $t > 0$, the set of all points X of K_0 for which $|X| \leqslant t$. Then the last inequality for Q shows that there exists a monotone increasing function $R(t)$ of t such that

$$\Lambda_r \text{ is } K_0^t\text{-admissible if } r \geqslant R(t). \qquad (13\cdot28)$$

Now the sphere $|X| \leqslant 1$ is obviously a subset of K_0, hence also of K_0^t if $t \geqslant 1$. Therefore, from (13·28),

$$|Q| \geqslant 1 \text{ for all points } Q \neq O \text{ of } \Lambda_r \text{ if } r \geqslant R(t) \text{ and } t \geqslant 1.$$

Also since
$$d(\Lambda_r) = d(\Lambda') \text{ for } r = 1, 2, 3, \ldots, \qquad (13\cdot29)$$

the sequence of lattices $\Lambda_1, \Lambda_2, \Lambda_3, \ldots$ is bounded.

But then, by theorem 2, this sequence contains a convergent infinite subsequence of lattices

$$\Lambda_{r_1}, \Lambda_{r_2}, \Lambda_{r_3}, \ldots,$$

say of limit Λ^*. Since, from (13·14),

$$d(\Lambda^*) = \lim_{k\to\infty} d(\Lambda_{r_k}) = d(\Lambda') < \Delta(K_0), \qquad (13\cdot30)$$

Λ^* cannot be K_0-admissible; there is then a point P^* of Λ^* which is an inner point of K_0 and so also an inner point of K_0^t if t is sufficiently large. Further, as in earlier proofs, it may be shown that there are points

$$P_{r_1}, P_{r_2}, P_{r_3}, \ldots \text{ of } \Lambda_{r_1}, \Lambda_{r_2}, \Lambda_{r_3}, \ldots \quad \text{respectively,}$$

such that

$$\lim_{k\to\infty} |P_{r_k} - P^*| = 0. \tag{13·31}$$

But then P_{r_k} is also an inner point of K_0^t if k is sufficiently large, contrary to (13·28). This completes the proof.

THEOREM 15. *There exists an unbounded star body of the finite type with a critical lattice which has no points on the boundary of this body.*

Proof. The same notation is used as in theorem 14, but it is assumed that $F(X)$ satisfies, instead of (13·2), the stronger condition

$$F(X) > F_0(X) \quad \text{if} \quad F_0(X) > 0; \tag{13·32}$$

e.g. take

$$F(X) = F_0(X)\left\{1 + \frac{F_0(X)}{|X|}\right\}. \tag{13·33}$$

Let Λ be a critical lattice of K_0. Since K is a subset of K_0, Λ is K-admissible; further, since from theorem 14,

$$d(\Lambda) = \Delta(K_0) = \Delta(K), \tag{13·34}$$

Λ is a critical lattice of K. But the boundary of K consists only of inner points of K_0, and so no point of Λ may lie on the boundary of K, as asserted.†

It is easily proved from § 15 that K_0 and so also K have an infinity of critical lattices. The question also arises:

PROBLEM B. *Do there exist critical lattices of K_0 which are not critical lattices of K, and do these lattices have points on the boundary of K?*

14. STAR BODIES WITH AUTOMORPHISMS

Let $X = (x_1, x_2, \ldots, x_n)$ and $X' = (x'_1, x'_2, \ldots, x'_n)$ be two points in R_n. The linear substitution

$$\Omega: \quad x'_g = \sum_{k=1}^{n} a_{gh} x_h, \quad \text{where} \quad g = 1, 2, \ldots, n, \tag{14·1}$$

of determinant

$$\omega = |a_{gh}|_{g,h=1,2,\ldots,n} \neq 0, \tag{14·2}$$

or shorter

$$X' = \Omega X \tag{14·3}$$

has an inverse

$$X = \Omega^{-1} X'. \tag{14·4}$$

The substitution defines a one-to-one mapping of R_n on itself.

---

† A much simpler proof of theorem 15 will be given in Part II of this paper.

If Λ is an arbitrary lattice, then $\Omega\Lambda$ denotes the lattice of all points $P' = \Omega P$ where P belongs to Λ; obviously

$$d(\Omega\Lambda) = |\omega|\, d(\Lambda). \tag{14·5}$$

THEOREM 16. *Let* K: $F(X) \leqslant 1$ *be a star body of the finite type,* Ω *a substitution of determinant* $\omega \neq 0$, $F'(X)$ *the distance function*

$$F'(X) = F(\Omega X), \tag{14·6}$$

and K' *the star body* $F'(X) \leqslant 1$. *Then* K' *is also of the finite type, and*

$$\Delta(K') = |\omega|^{-1} \Delta(K). \tag{14·7}$$

Proof. If Λ is any K-admissible lattice, then $\Lambda' = \Omega^{-1}\Lambda$ is evidently K'-admissible, and so K' is also of the finite type. Further $\Delta(K')$ is not greater than the lower bound of $d(\Omega^{-1}\Lambda) = |\omega|^{-1} d(\Lambda)$ extended over all K-admissible lattices, i.e.

$$\Delta(K') \leqslant |\omega|^{-1} \Delta(K). \tag{14·8}$$

Since $F(X) = F'(\Omega^{-1}X)$, conversely

$$\Delta(K) \leqslant |\omega|\, \Delta(K'). \tag{14·9}$$

From these two inequalities, the assertion follows at once.

DEFINITION 8. *The linear substitution* $X' = \Omega X$ *is called an automorphism of the star body* K: $F(X) \leqslant 1$, *if identically in* X,

$$F(X') = F(X). \tag{14·10}$$

It is obvious that such an automorphism leaves both K and its boundary C invariant.

THEOREM 17. *If the star body* K *is of the finite type and admits the automorphism* $X' = \Omega X$ *of determinant* ω, *then* $\omega = \pm 1$.

Proof. By theorem 16, $\Delta(K) = |\omega|^{-1} \Delta(K)$, whence $|\omega| = 1$ since $\Delta(K) \neq 0$.

This theorem shows that star bodies having automorphisms of determinant $\omega \neq \pm 1$, are necessarily of the infinite type, e.g. the star body of distance function $F(X) = |x_1^2 x_2 \ldots x_n|^{1/(n+1)}$ with the automorphism

$$x_1' = t^{-\frac{1}{2}(n-1)} x_1,\ x_2' = t x_2,\ \ldots,\ x_n' = t x_n \quad (t > 0). \tag{14·11}$$

It is obvious that if K is of the finite type, then the set of all automorphisms of K forms a group. Whether this group is finite or infinite, discrete or continuous, depends on K itself.

DEFINITION 9. *An unbounded star body* K *of the finite type is called automorphic if it admits a group* Γ *of automorphisms* Ω *with the following property: 'There exists a positive constant* c *depending only on* K *and* Γ *such that to every point* X *of* K *there is an element* Ω *of* Γ *satisfying*

$$|\Omega X| \leqslant c.\text{'} \tag{14·12}$$

A few examples of automorphic star bodies are given in the next section.

15. EXAMPLES OF AUTOMORPHIC STAR BODIES

(1) Let $r \geqslant 0$ and $s \geqslant 0$ be integers such that $r+2s = n$, and let $F(X)$ be the distance function

$$F(X) = \left| \prod_{\rho=1}^{r} x_\rho \prod_{\sigma=1}^{s} (x_{r+\sigma}^2 + x_{r+s+\sigma}^2) \right|^{1/n}. \tag{15·1}$$

It was shown in the first part of the proof of theorem 5 that the star body $K: F(X) \leqslant 1$ is of the finite type if $r = n, s = 0$. Just the same proof applies when $s > 0$, except that the field $\mathfrak{K}$ there must now be algebraic with r real and $2s$ complex conjugate fields. If the trivial cases $r = 1$, $s = 0$ and $r = 0$, $s = 1$ be excluded, then K is not bounded and admits a continuous group of automorphisms depending on $n-1$ parameters, namely, the group of substitutions

$$x'_\rho = t_\rho x_\rho, \quad \text{where} \quad \rho = 1, 2, \ldots, r, \tag{15·2}$$

$$x'_{r+\sigma} = t_{r+\sigma} x_{r+\sigma} - t_{r+s+\sigma} x_{r+s+\sigma}, \quad x'_{r+s+\sigma} = t_{r+s+\sigma} x_{r+\sigma} + t_{r+\sigma} x_{r+s+\sigma},$$

where $$\sigma = 1, 2, \ldots, s, \tag{15·3}$$

while $t_1, t_2, \ldots, t_n$ are n real numbers such that

$$\prod_{\rho=1}^{r} t_\rho \prod_{\sigma=1}^{s} (t_{r+\sigma}^2 + t_{r+s+\sigma}^2) = \pm 1. \tag{15·4}$$

The star body K is automorphic since obviously every point X of K can be transformed into a point X' of bounded distance from O by one of these automorphisms.

(2) Let r be an integer such that $1 \leqslant r \leqslant n-1$, and let K be the star body of distance function

$$F(X) = \left| \sum_{\rho=1}^{r} x_\rho^2 - \sum_{\sigma=r+1}^{n} x_\sigma^2 \right|^{\frac{1}{2}}. \tag{15·5}$$

By the theory of quadratic forms, K admits a group of automorphisms depending on $\frac{1}{2}n(n-1)$ real parameters. It is again possible to show that every point in K can be transformed by one of these automorphisms into a point of bounded distance from O. Hence K is automorphic provided it is of the finite type, and so the following problem arises:

PROBLEM C. *Is the star body of distance function*

$$F(X) = \left| \sum_{\rho=1}^{r} x_\rho^2 - \sum_{\sigma=r+1}^{n} x_\sigma^2 \right|^{\frac{1}{2}} \tag{15·6}$$

of the finite or of the infinite type?[†]

For $2 \leqslant n \leqslant 4$, K is of the finite type, because there exist indefinite quadratic forms in n variables with integral coefficients and of given signature which do not represent zero non-trivially. If, however, $n \geqslant 5$, then, by Meyer's theorem (Bachmann 1898), every indefinite quadratic form with integral coefficients does represent zero; so the solution of problem C may be difficult.

[†] *Addition, May* 1946. In a joint paper, H. Davenport and H. Heilbron have just shown that K is of the infinite type if $n \geqslant 5$.

(3) Let $n = 2$, and denote by θ any number with $0 < \theta < 1$. The line segments joining the pairs of points

$$(\theta^k, \theta^{-k}) \quad \text{and} \quad (\theta^{k+1}, \theta^{-k-1}), \quad \text{where} \quad k = 0, \pm 1, \pm 2, \ldots,$$

form an infinite polygon Π; let C be the curve consisting of Π and the images of Π in O and the two axes. Then C forms the complete boundary of a two-dimensional star body K. There is no difficulty in proving that K is of the finite type and that it admits the infinite group of automorphisms

$$x_1' = \pm \theta^k x_g, \quad x_2' = \pm \theta^{-k} x_h, \quad \text{where} \quad k = 0, \pm 1, \pm 2, \ldots$$

and

$$g = 1, h = 2 \quad \text{or} \quad g = 2, h = 1. \tag{15·7}$$

It can be shown that every point of K can be transformed by one of these automorphisms into a point of bounded distance from O; hence K is an automorphic star body.

16. PROPERTIES OF THE LATTICE FUNCTION $F(\Lambda)$

It was seen in § 11 that $F(\Lambda)$ need not be a continuous function of Λ. The next two theorems on sequences of lattices have therefore some interest:

THEOREM 18. *Let $\Lambda_1, \Lambda_2, \Lambda_3, \ldots$ be a convergent sequence of lattices, say of limit Λ. Then*

$$F(\Lambda) \geqslant \liminf_{r\to\infty} F(\Lambda_r). \tag{16·1}$$

Proof. Choose reduced bases $Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_n^{(r)}$ of Λ_r, and a basis $Y_1, Y_2, \ldots, Y_n$ of Λ such that

$$\lim_{r\to\infty} | Y_g^{(r)} - Y_g | = 0, \quad \text{where} \quad g = 1, 2, \ldots, n. \tag{16·2}$$

Every point $P \neq O$ of Λ can be written as

$$P = u_1 Y_1 + \ldots + u_n Y_n \tag{16·3}$$

with integral coefficients $u_1, \ldots, u_n$ not all zero. Put

$$P_r = u_1 Y_1^{(r)} + \ldots + u_n Y_n^{(r)}, \tag{16·4}$$

then

$$P_r \neq O, \text{ and } P_r \text{ lies in } \Lambda_r. \tag{16·5}$$

Hence

$$F(P_r) \geqslant F(\Lambda_r). \tag{16·6}$$

Therefore by the continuity of $F(X)$,

$$F(P) = \lim_{r\to\infty} F(P_r) \geqslant \liminf_{r\to\infty} F(\Lambda_r), \tag{16·7}$$

as asserted.

THEOREM 19. *Let $\Lambda_1, \Lambda_2, \Lambda_3, \ldots$ be a convergent sequence of lattices, say of limit Λ, and assume that $\phi = \lim_{r\to\infty} F(\Lambda_r)$ exists and is positive. Let there also be a constant $c > 0$ and an infinite sequence of points $P_1, P_2, P_3, \ldots$ such that*

$$P_r \neq O; \quad |P_r| \leqslant c; \quad P_r \text{ lies in } \Lambda_r, \quad \text{where} \quad r = 1, 2, 3, \ldots, \tag{16·8}$$

$$\lim_{r\to\infty} F(P_r) \textit{ exists and is equal to } \phi.$$

Then
$$\lim_{r\to\infty} F(\Lambda_r) = F(\Lambda), \tag{16·9}$$
and there exists a point $P \neq O$ *of* Λ *such that*
$$F(P) = F(\Lambda). \tag{16·10}$$

Proof. There is a positive number ρ such that the sphere $|X| \leqslant \rho$ is contained in the star body F: $F(X) \leqslant 1$. Put
$$\sigma = \tfrac{1}{2}\rho\phi. \tag{16·11}$$
Then the sphere $|X| \leqslant \sigma$ is contained in the star body $F(X) \leqslant \sigma/\rho$, i.e. in $F(X) \leqslant F(\Lambda_r)$, for all sufficiently large r, say for $r \geqslant r_0$. Therefore for every point $Q \neq O$ of Λ_r, since $F(Q) \geqslant F(\Lambda_r)$,
$$|Q| \geqslant \sigma \quad \text{if} \quad r \geqslant r_0. \tag{16·12}$$
Let, in particular, $Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_n^{(r)}$ be a reduced basis of Λ_r and $Y_1, Y_2, \ldots, Y_n$ a basis of Λ taken such that
$$\lim_{r\to\infty} |Y_g^{(r)} - Y_g| = 0, \quad \text{where} \quad g = 1, 2, \ldots, n. \tag{16·13}$$
Then
$$|Y_g^{(r)}| \geqslant \sigma \quad \text{for} \quad r \geqslant r_0, \quad g = 1, 2, \ldots, n. \tag{16·14}$$

On the other hand, from theorem 1,
$$|Y_1^{(r)}|\,|Y_2^{(r)}| \ldots |Y_n^{(r)}| \leqslant \gamma_n d(\Lambda_r). \tag{16·15}$$
Also, from the hypothesis,
$$\lim_{r\to\infty} d(\Lambda_r) = d(\Lambda), \tag{16·16}$$
hence
$$\tfrac{1}{2}d(\Lambda) \leqslant d(\Lambda_r) \leqslant 2d(\Lambda) \quad \text{for} \quad r \geqslant r_1, \text{ say}, \tag{16·17}$$
and so
$$|Y_g^{(r)}| \leqslant 2\sigma^{-(n-1)}\gamma_n d(\Lambda) \quad \text{for} \quad r \geqslant \max(r_0, r_1), \quad \text{where} \quad g = 1, 2, \ldots, n. \tag{16·18}$$

Since P_r is a point of Λ_r different from O,
$$P_r = u_1^{(r)} Y_1^{(r)} + \ldots + u_n^{(r)} Y_n^{(r)} \tag{16·19}$$
with integral coefficients $u_1^{(r)}, \ldots, u_n^{(r)}$ not all zero. On solving this vector equation for $u_1^{(r)}, \ldots, u_n^{(r)}$,
$$d(\Lambda_r)\,|u_g^{(r)}| = |\{Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_n^{(r)}\}|\,|u_g^{(r)}| = |\{Y_1^{(r)}, \ldots, Y_{g-1}^{(r)}, P_r, Y_{g+1}^{(r)}, \ldots, Y_n^{(r)}\}|. \tag{16·20}$$
Hence the lower bound for $d(\Lambda_r)$ and the upper bounds for $Y_g^{(r)}$ and P_r imply that
$$|u_g^{(r)}| \leqslant c', \tag{16·21}$$
where c' is a positive number independent of r and g.

There exists then an infinite sequence of indices
$$r = r_1, r_2, r_3, \ldots, \quad \text{where} \quad \lim_{k\to\infty} r_k = \infty, \tag{16·22}$$
such that the coefficients
$$u_g^{(r_k)} = u_g \text{ say}, \quad \text{where} \quad k = 1, 2, 3, \ldots;\ g = 1, 2, \ldots, n, \tag{16·23}$$

assume integral values independent of k, and such that at least one of these integers $u_1, \ldots, u_n$ is different from zero. Further

$$P_{rk} = u_1 Y_1^{(rk)} + \ldots + u_n Y_n^{(rk)}, \quad \text{where} \quad k = 1, 2, 3, \ldots, \tag{16·24}$$

and so the points P_{rk} tend to the limit point

$$P = u_1 Y_1 + \ldots + u_n Y_n \neq O \tag{16·25}$$

which is a point of Λ: From the hypothesis

$$F(P) = \lim_{k\to\infty} F(P_{rk}) = \lim_{r\to\infty} F(P_r) = \lim_{r\to\infty} F(\Lambda_r), \tag{16·26}$$

whence $$F(\Lambda) \leqslant \lim_{r\to\infty} F(\Lambda_r). \tag{16·27}$$

Moreover, from the last theorem,

$$F(\Lambda) \geqslant \lim_{r\to\infty} F(\Lambda_r), \tag{16·28}$$

and so the assertion follows at once.

17. LATTICE POINTS ON THE BOUNDARY OF AN AUTOMORPHIC STAR BODY

THEOREM 20. *Let K: $F(X) \leqslant 1$ be an automorphic star body, and let Λ be any lattice such that $F(\Lambda) > 0$. Then there exists a lattice Λ^* and a point P^* of Λ^* such that*

$$F(P^*) = F(\Lambda^*) = F(\Lambda), \quad d(\Lambda^*) = d(\Lambda). \tag{17·1}$$

(*Remark.* Λ^* need not be different from Λ. The theorem remains valid if $F(\Lambda) = 0$, but then is nearly trivial.)

Proof. Assume that Λ contains no point P such that

$$F(P) = F(\Lambda); \tag{17·2}$$

otherwise the assertion is certainly true. There exists then an infinite sequence of points $P_1, P_2, P_3 \ldots$ of Λ such that

$$\lim_{r\to\infty} F(P_r) = F(\Lambda) > 0; \tag{17·3}$$

assume that all these points are different from O.

For each point P_r select an automorphism Ω_r of K such that

$$|\Omega_r P_r| \leqslant c. \tag{17·4}$$

Put $$\Omega_r P_r = Q_r, \quad \Omega_r \Lambda = \Lambda_r, \tag{17·5}$$

so that Q_r belongs to Λ_r, is different from O, and satisfies the inequality

$$|Q_r| \leqslant c. \tag{17·6}$$

By the invariance of K, $$F(Q_r) = F(\Omega_r P_r) = F(P_r), \tag{17·7}$$

hence from the hypothesis $$\lim_{r\to\infty} F(Q_r) = F(\Lambda) > 0. \tag{17·8}$$

Further, from theorem 17, Ω_r is of determinant ± 1, and so

$$d(\Lambda_r) = d(\Lambda). \tag{17·9}$$

Next, it is shown that $$F(\Lambda_r) = F(\Lambda). \tag{17·10}$$

For if P runs over all points of $\Lambda - [O]$, then $Q = \Omega_r P$ runs over all points of $\Lambda_r - [O]$, and vice versa. But by the invariance assumption,

$$F(Q) = F(P), \tag{17·11}$$

and by definition,

$$F(\Lambda) = \underset{P \text{ in } \Lambda - [O]}{\text{l.b.}} F(P), \quad F(\Lambda_r) = \underset{Q \text{ in } \Lambda_r - [O]}{\text{l.b.}} F(Q), \tag{17·12}$$

whence (17·10) follows at once.

Finally, the sequence of lattices

$$\Lambda_1, \Lambda_2, \Lambda_3, \ldots$$

is bounded. For from (17·9), the determinants $d(\Lambda_r)$ are bounded, and from (17·10),

$$F(Q) \geqslant F(\Lambda) \text{ for all points } Q \neq O \text{ of } \Lambda_r. \tag{17·13}$$

Hence, if ρ is any number such that K contains the sphere $|X| \leqslant \rho$, i.e. $F(\Lambda)K$ contains the sphere $|X| \leqslant \rho F(\Lambda)$, then

$$|Q| \geqslant F(\Lambda)\rho \text{ for all points } Q \neq O \text{ of } \Lambda_r. \tag{17·14}$$

From theorem 2, there exists then an infinite subsequence of lattices

$$\Lambda_{r_1}, \Lambda_{r_2}, \Lambda_{r_3}, \ldots$$

which tends to a limit, say the lattice Λ^*; from (17·9)

$$d(\Lambda^*) = \lim_{k \to \infty} d(\Lambda_{r_k}) = d(\Lambda). \tag{17·15}$$

Hence the supposition of theorem 19 is satisfied if one substitutes therein for the sequence of lattices $\{\Lambda_r\}$, the lattice Λ, and the sequence of points $\{P_r\}$ respectively, the sequence of lattices $\{\Lambda_{r_k}\}$, the lattice Λ^*, and the sequence of points $\{Q_{r_k}\}$ of the present proof. The assertion follows therefore at once from theorem 19.

Remark. Theorem 20 does not assert that every lattice Λ^* satisfying

$$F(\Lambda^*) = F(\Lambda), \quad d(\Lambda^*) = d(\Lambda) \tag{17·16}$$

contains a point P^* such that $F(P^*) = F(\Lambda^*)$. Thus take $n = 2$ and $F(X) = |x_1 x_2|^{\frac{1}{2}}$. Then, as follows from results in the theory of indefinite binary quadratic forms (Koksma 1936), there exists an infinity of lattices Λ^* such that

$$F(\Lambda^*) = 1, \quad d(\Lambda^*) = 3, \tag{17·17}$$

and some, but not all, of these lattices contain points P^* such that

$$F(P^*) = 1. \tag{17·18}$$

The following particular case of the last theorem is of special interest.

THEOREM 21. *Every automorphic star body K has a critical lattice with at least one point on the boundary of K.*

Proof. A lattice Λ is a critical lattice of K if and only if

$$F(\Lambda) = 1, \quad d(\Lambda) = \Delta(K). \tag{17·19}$$

Now, from theorem 8, critical lattices of K do exist; the assertion follows therefore at once from theorem 20.

PROBLEM D. *Does* every *critical lattice of an automorphic star body K have at least one point on the boundary of K?*

The example in theorem 20 does not answer this question, but makes it probable that the answer is in the negative.

Theorem 20 further suggests the following:

PROBLEM E. *To study the set d_F of the values of $d(\Lambda)$ where Λ runs over all lattices Λ satisfying $F(\Lambda) = 1$.*

The set d_F has a smallest element which is, of course, $\Delta(K)$; this number and the other elements of the set may be considered as the successive minima of the lattice point problem for the body K: $F(X) \leqslant 1$. Even in the case $F(X) = |x_1x_2|^{\frac{1}{2}}$, d_F is a very complicated set (Koksma 1936), and the same is to be expected for other unbounded star bodies. It is then rather surprising that in the case of automorphic star bodies, all these minima are actually attained in the sense that to every element δ of d_F there exists a lattice Λ^* and a point P^* of Λ^* such that

$$F(P^*) = F(\Lambda^*) = 1, \quad d(\Lambda^*) = \delta. \tag{17·20}$$

18. THE INVARIANT SUBSET OF AN AUTOMORPHIC STAR BODY

Let K: $F(X) \leqslant 1$ be an automorphic star body, and let Γ be a group of automorphisms Ω of K. We denote by Σ_Γ the set of the points X in R_n which have the following property:

'There exists a positive number $a(X)$ depending only on X such that

$$|\Omega X| \leqslant a(X) \text{ for all } \Omega \text{ in } \Gamma.\text{'} \tag{18·1}$$

This set Σ_Γ is called the invariant manifold of K. It may contain only the origin, and it has the following properties:

(*a*) If X lies in Σ_Γ, and Ω is an element of Γ, then $Y = \Omega X$ also lies in Σ_Γ, and we may take

$$a(Y) = a(X). \tag{18·2}$$

For let Ω_1 be an arbitrary element of Γ. Then $\Omega_2 = \Omega_1\Omega$ also belongs to Γ, and so by the definition of $a(X)$,

$$|\Omega_1 Y| = |\Omega_2 X| \leqslant a(X). \tag{18·3}$$

(*b*) If $X_1, X_2, \ldots, X_m$ is any number of points of Σ_Γ, and if $t_1, t_2, \ldots, t_m$ are real numbers, then $t_1X_1+t_2X_2+\ldots+t_mX_m$ also lies in Σ_Γ, and we may take

$$a(t_1X_1+t_2X_2+\ldots+t_mX_m) = |\,t_1\,|\,a(X_1)+|\,t_2\,|\,a(X)+\ldots+|\,t_m\,|\,a(X_m). \quad (18{\cdot}4)$$

For if Ω is any element of Γ, then

$$\begin{aligned} |\,\Omega(t_1X_1+\ldots+t_mX_m)\,| &= |\,t_1\Omega X_1+\ldots+t_m\Omega X_m\,| \\ &\leqslant |\,t_1\,|\,|\,\Omega X_1\,|+\ldots+|\,t_m\,|\,|\,\Omega X_m\,| \leqslant |\,t_1\,|\,a(X_1)+\ldots+|\,t_m\,|\,a(X_m). \quad (18{\cdot}5) \end{aligned}$$

From (*b*), Σ_Γ is a linear manifold. Let it be of dimension δ where $0 \leqslant \delta \leqslant n$, and let $P_1, \ldots, P_\delta$ be a set of δ independent points of Σ_Γ. Then the points X of Σ_Γ may be written as

$$X = \xi_1P_1+\ldots+\xi_\delta P_\delta \quad (18{\cdot}6)$$

with real coefficients $\xi_1, \ldots, \xi_\delta$; conversely, every such point X belongs to Σ_δ. On considering this vector equation as a system of n equations for the n co-ordinates, we find on solving for $\xi_1, \ldots, \xi_\delta$ that

$$\max(|\,\xi_1\,|, \ldots, |\,\xi_\delta\,|) \leqslant \gamma|\,X\,|, \quad (18{\cdot}7)$$

where γ is a positive number depending only on the choice of $P_1, \ldots, P_\delta$.

(*c*) There exists a positive constant b such that if X is any point of Σ_Γ, Ω any element of Γ, and $Y = \Omega X$, then

$$b^{-1}|\,X\,| \leqslant |\,Y\,| \leqslant b|\,X\,|. \quad (18{\cdot}8)$$

For let $X = \xi_1P_1+\ldots+\xi_\delta P_\delta$. Then

$$\begin{aligned} |\,Y\,| = |\,\xi_1\Omega P_1+\ldots+\xi_\delta\Omega P_\delta\,| &\leqslant \max(|\,\xi_1\,|, \ldots, |\,\xi_\delta\,|)\,(|\,\Omega P_1\,|+\ldots+|\,\Omega P_\delta\,|) \\ &\leqslant \gamma|\,X\,|\,\{a(P_1)+\ldots+a(P_\delta)\} = b|\,X\,|, \quad (18{\cdot}9) \end{aligned}$$

where $$b = \gamma\{a(P_1)+\ldots+a(P_\delta)\}.$$

Further if X is in Σ_Γ and $Y = \Omega X$, then Y is also in Σ_Γ and $X = \Omega^{-1}Y$. Hence by the same proof $|\,X\,| \leqslant b|\,Y\,|$, whence the assertion.

Let now $J_\Gamma = K \times \Sigma_\Gamma$ be the set of all points of Σ_Γ which belong to K; we call J_Γ the invariant subset of K.

(*d*) The invariant subset J_Γ is a bounded set. For let X be any point of J_Γ. By definition 9, there exists a positive constant c and an element Ω of Γ such that

$$|\,\Omega X\,| \leqslant c. \quad (18{\cdot}10)$$

Hence from (*c*), $$|\,X\,| \leqslant b|\,\Omega X\,| \leqslant bc, \quad (18{\cdot}11)$$

as asserted.

This result shows that the dimension δ of Σ_Γ and J_Γ is at most $n-1$. For let this assertion be false so that $\delta = n$. Then Σ_Γ coincides with the whole space R_n, and therefore J_Γ is identical with K. Hence, from (*d*), K is a bounded set, contrary to the definition of an automorphic star body.

Probably δ satisfies the stronger inequality $\delta \leqslant n-2$. The following example shows, however, that δ can be any integer in the interval

$$0 \leqslant \delta \leqslant n-2.$$

Take for K the star body of distance function

$$F(X) = \max\left(\{x_1^2+\ldots+x_\delta^2\}^{\frac{1}{2}}, \, |x_{\delta+1}\ldots x_n|^{1/(n-\delta)}\right), \tag{18·12}$$

and for Γ the group of automorphisms

$$x_1' = x_1, \ldots, x_\delta' = x_\delta, \quad x_{\delta+1}' = t_{\delta+1}x_{\delta+1}, \ldots, x_n' = t_n x_n, \tag{18·13}$$

where $t_{\delta+1}, \ldots, t_n$ are real numbers of product $t_{\delta+1}\ldots t_n = 1$; then Σ_Γ is the δ-dimensional linear manifold

$$x_{\delta+1} = \ldots = x_n = 0. \tag{18·14}$$

The automorphic star bodies with $\delta = 0$ are of particular interest; then both Σ_Γ and J_Γ reduce to the single point O. To this type belong, for instance, all the star bodies considered in § 15. In § 20, a general property of star bodies with $\delta = 0$ will be proved.

19. AN IMPROVEMENT ON THEOREM 13

THEOREM 22. *Let K: $F(X) \leqslant 1$ be any star body of the finite type. Then there exists to every number $\epsilon > 0$ a positive number $t = t(\epsilon)$ such that every critical lattice Λ of K contains at least one point P satisfying the inequalities*

$$1 \leqslant F(P) < 1+\epsilon, \quad |P| \leqslant t. \tag{19·1}$$

Proof. By the remark to theorem 10, there is a positive number $t^* = t^*(\epsilon)$ such that the star body

$$K^* = K^{(t^*)}: \qquad F(X) \leqslant 1, \quad |X| \leqslant t^*$$

is of determinant

$$\Delta(K^*) \geqslant \left(1+\frac{\epsilon}{2}\right)^{-n} \Delta(K). \tag{19·2}$$

Put

$$t = \left(1+\frac{\epsilon}{2}\right)t^*, \quad K^{**} = \left(1+\frac{\epsilon}{2}\right)K^*, \tag{19·3}$$

so that K^{**} consists of the points satisfying

$$F(X) \leqslant 1+\frac{\epsilon}{2}, \quad |X| \leqslant \left(1+\frac{\epsilon}{2}\right)t^* = t, \tag{19·4}$$

then

$$\Delta(K^{**}) = \left(1+\frac{\epsilon}{2}\right)^n \Delta(K^*) \geqslant \Delta(K). \tag{19·5}$$

Hence every lattice of determinant $\Delta(K)$ contains a point $P \neq O$ for which

$$F(P) \leqslant 1+\frac{\epsilon}{2} < 1+\epsilon, \quad |P| \leqslant t; \tag{19·6}$$

if the lattice is critical with respect to K, then moreover

$$F(P) \geqslant 1, \tag{19·7}$$

whence the assertion.

20. AUTOMORPHIC STAR BODIES WITH $\Sigma_\Gamma = J_\Gamma = \{O\}$

THEOREM 23. *Let K: $F(X) \leqslant 1$ be an automorphic star body for which Σ_Γ and so also J_Γ consist of the single point O. Further let Λ be any critical lattice of K, and ϵ any positive number. Then there exists an infinite sequence of different points $P_1, P_2, P_3, \ldots$ of Λ such that*

$$1 \leqslant F(P_\mu) < 1+\epsilon, \quad \textit{where} \quad \mu = 1, 2, 3, \ldots. \tag{20·1}$$

Proof. Assume the assertion is false. There is then a positive number ϵ and a critical lattice Λ of K such that the inequality

$$1 \leqslant F(P_\mu) < 1+\epsilon \tag{20·2}$$

is satisfied by only a finite number of points of Λ, say by only the m points

$$P_1,\ P_2,\ \ldots,\ P_m;$$

by the last theorem, m is not zero. It may be assumed, without loss of generality, that ϵ and Λ have been chosen so as to make m a minimum, that is,

> There does not exist any positive number ϵ^* and any critical lattice Λ^* of K such that the inequality
>
> $$1 \leqslant F(P^*_\mu) < 1+\epsilon^* \tag{20·3}$$
>
> is satisfied by less than m points P^*_μ of Λ^*.

This minimum assumption implies, in particular, that

$$F(P_\mu) = 1, \quad \text{where} \quad \mu = 1, 2, \ldots, m; \tag{20·4}$$

for if, for instance, $F(P_m) = 1+\delta > 1$, then, on putting $\epsilon^* = \delta$, $\Lambda^* = \Lambda$, there are less than m points P^* of Λ^* such that

$$1 \leqslant F(P^*) < 1+\epsilon^*. \tag{20·5}$$

Let now Ω be any automorphism in Γ. Then from (20·2), (20·4) and theorem 22, the lattice $\Omega\Lambda$ has the following properties:

> There are just m points P^* of $\Omega\Lambda$ for which
>
> $$1 \leqslant F(P^*) < 1+\epsilon, \tag{20·6}$$
>
> viz. the points $$P^* = \Omega P_1,\ \Omega P_2,\ \ldots,\ \Omega P_m; \tag{20·7}$$
>
> and, in fact, $$F(\Omega P_\mu) = 1, \quad \text{where} \quad \mu = 1, 2, \ldots, m. \tag{20·8}$$
>
> There is, moreover, a positive number t independent of Ω and μ such that
>
> $$|\Omega P_\mu| \leqslant t \text{ for at least one index } \mu \text{ with } 1 \leqslant \mu \leqslant m. \tag{20·8½}$$

From (20·4), P_m is different from O, and so does not belong to Σ_Γ. Hence there exists an infinite sequence

$$\{\Omega_r^{(m)}\} = \{\Omega_1^{(m)}, \Omega_2^{(m)}, \Omega_3^{(m)}, \ldots\} \tag{20·9}$$

of automorphisms $\Omega_r^{(m)}$ of K such that

$$\lim_{r\to\infty} |\Omega_r^{(m)} P_m| = \infty. \tag{20·10}$$

Now construct $m-1$ infinite subsequences

$$\{\Omega_r^{(\mu)}\} = \{\Omega_1^{(\mu)}, \Omega_2^{(\mu)}, \Omega_3^{(\mu)}, \ldots\} \tag{20·11}$$

of $\{\Omega_r^{(m)}\}$ according to the following rule:

Suppose the sequence $\{\Omega_r^{(\mu+1)}\}$ has been defined. If now

$$\lim_{r\to\infty} |\Omega_r^{(\mu+1)} P_\mu| = \infty, \tag{20·12}$$

then let $\{\Omega_r^{(\mu)}\}$ be identical with $\{\Omega_r^{(\mu+1)}\}$:

$$\Omega_r^{(\mu)} = \Omega_r^{(\mu+1)}, \quad \text{where} \quad r = 1, 2, 3, \ldots. \tag{20·13}$$

If, however,

$$\liminf_{r\to\infty} |\Omega_r^{(\mu+1)} P_\mu| \tag{20·14}$$

is finite, then choose for $\{\Omega_r^{(\mu)}\}$ an infinite subsequence

$$\Omega_1^{(\mu)} = \Omega_{r_1}^{(\mu+1)}, \ \Omega_2^{(\mu)} = \Omega_{r_2}^{(\mu+1)}, \ \Omega_3^{(\mu)} = \Omega_{r_3}^{(\mu+1)}, \ \ldots \tag{20·15}$$

of $\{\Omega_r^{(\mu+1)}\}$ such that the point sequence

$$\{\Omega_1^{(\mu)} P_\mu, \ \Omega_2^{(\mu)} P_\mu, \ \Omega_3^{(\mu)} P_\mu, \ \ldots\} \tag{20·16}$$

tends to a limit, say the point P_μ^*.

This means that the last sequence $\{\Omega_r^{(1)}\}$ has the following properties:

$$\lim_{r\to\infty} |\Omega_r^{(1)} P_m| = \infty. \tag{20·17}$$

If μ is one of the indices $1, 2, \ldots, m-1$, then either

$$\lim_{r\to\infty} |\Omega_r^{(1)} P_\mu| = \infty, \tag{20·18}$$

or there exists a finite point P_μ^* such that

$$\lim_{r\to\infty} |\Omega_r^{(1)} P_\mu - P_\mu^*| = 0. \tag{20·19}$$

Denote then by $\mu_1, \mu_2, \ldots, \mu_g$ those different indices μ with $1 \leqslant \mu \leqslant m$ for which

$$\lim_{r\to\infty} |\Omega_r^{(1)} P_\mu| = \infty, \tag{20·20}$$

by $\mu_1^*, \mu_2^*, \ldots, \mu_h^*$ those for which

$$\lim_{r\to\infty} |\Omega_r^{(1)} P_\mu - P_\mu^*| = 0; \tag{20·21}$$

hence $g + h = m$. From (20·11) and (20·17), then

$$g \geqslant 1, \quad h \geqslant 1. \tag{20·22}$$

Since $\Omega_r^{(1)}$ is an automorphism of K, it is evident that

$$d(\Omega_r^{(1)}\Lambda) = d(\Lambda) = \Delta(K), \quad F(\Omega_r^{(1)}\Lambda) = F(\Lambda) = 1, \quad \text{where} \quad r = 1, 2, 3, \ldots \tag{20·23}$$

(for the second equation, compare the proof of theorem 20), and so the lattices

$$\{\Omega_1^{(1)}\Lambda, \Omega_2^{(1)}\Lambda, \Omega_3^{(1)}\Lambda, \ldots\} \tag{20·24}$$

form a bounded sequence. From theorem 2, one can therefore choose an infinite subsequence $\{\Omega_r\}$ of automorphisms

$$\Omega_1 = \Omega_{r'}^{(1)}, \Omega_2 = \Omega_{r''}^{(1)}, \Omega_3 = \Omega_{r'''}^{(1)} \ldots \tag{20·25}$$

in $\{\Omega_r^{(1)}\}$ such that the corresponding sequence of lattices

$$\Lambda_1 = \Omega_1\Lambda, \Lambda_2 = \Omega_2\Lambda, \Lambda_3 = \Omega_3\Lambda, \ldots \tag{20·26}$$

tends to a limit, the lattice Λ^*, say.

Then from (20·23) it follows that

$$\lim_{r\to\infty} d(\Lambda_r) = \Delta(K), \quad \lim_{r\to\infty} F(\Lambda_r) = 1. \tag{20·27}$$

Further, from (20·8½), and the construction of $\{\Omega_r^{(\mu)}\}$ and $\{\Lambda_r\}$, each lattice

$$\Lambda_r = \Omega_r\Lambda, \quad \text{where} \quad r = 1, 2, 3, \ldots, \tag{20·28}$$

contains a point $P^{(r)} = \Omega_r P_{\mu(r)}$ with $1 \leqslant \mu(r) \leqslant m-1$, (20·29)

such that $P^{(r)} \neq O, \quad |P^{(r)}| \leqslant t, \quad F(P^{(r)}) = 1.$ (20·30)

An application of theorem 19 therefore gives

$$d(\Lambda^*) = \lim_{r\to\infty} d(\Lambda_r) = \Delta(K), \quad F(\Lambda^*) = \lim_{r\to\infty} F(\Lambda_r) = 1, \tag{20·31}$$

which means that Λ^* is a critical lattice. Now a consideration analogous to that in earlier proofs makes it evident that the points

$$P^*_{\mu_1^*}, P^*_{\mu_2^*}, \ldots, P^*_{\mu_h^*}$$

as defined in (20·21), are the only points P^* of Λ^* such that

$$1 \leqslant F(P^*) < 1 + \frac{\epsilon}{2}; \tag{20·32}$$

moreover $F(P^*_{\mu_1^*}) = F(P^*_{\mu_2^*}) = \ldots = F(P^*_{\mu_h^*}) = 1.$ (20·33)

Hence Λ^* is a lattice of the same type as Λ, except that m is replaced by the smaller number h. This contradicts the minimum assumption (20·3); the hypothesis is therefore false and the assertion is true.

PROBLEM F. *Does the assertion of theorem 23 remain true if Σ_Γ is of positive dimension δ?*

Closely related to problem F is the following question which I also have not been able to solve:

PROBLEM G. *To decide whether there exists an automorphic star body K: $F(X) \leqslant 1$ with the following two properties: (a) The invariant manifold Σ_Γ is of positive dimension. (b) There exists a critical lattice Λ of K and a positive number α such that*

$$F(P) \geqslant 1+\alpha \tag{20·34}$$

for all points P of Λ which do not belong to Σ_Γ.

21. STAR BODIES OF RANK δ

The considerations in § 18 can be generalized and lead to the following definition:

DEFINITION 10. *Let K: $F(X) \leqslant 1$ be a star body of the finite type with a group Γ of automorphisms Ω, and let δ be an integer such that $1 \leqslant \delta \leqslant n-1$. Then K is said to be of rank δ with respect to Γ if δ is the largest integer such that to every positive number t^* and to every δ-dimensional linear manifold M containing O there is an element $\Omega = \Omega(t^*, M)$ of Γ satisfying*

$$|\Omega X| \geqslant t^* F(X) \quad \text{for all points } X \text{ of } M. \tag{21·1}$$

An example on this definition is given by

THEOREM 24. *Let K be the star body of distance function*

$$F(X) = \left| \prod_{\rho=1}^{r} x_\rho \prod_{\sigma=1}^{s} (x_{r+\sigma}^2 + x_{r+s+\sigma}^2) \right|^{1/n}, \quad \text{where} \quad r+2s = n, \tag{21·2}$$

and let Γ be the group of all automorphisms Ω of K defined by

$$\Omega: \begin{cases} x'_\rho = t_\rho x_\rho, & \text{where} \quad \rho = 1, 2, \dots, r, \quad (21·3) \\ \left.\begin{aligned} x'_{r+\sigma} &= t_{r+\sigma} x_{r+\sigma} - t_{r+s+\sigma} x_{r+s+\sigma} \\ x'_{r+s+\sigma} &= t_{r+s+\sigma} x_{r+\sigma} + t_{r+\sigma} x_{r+s+\sigma} \end{aligned}\right\} & \text{where} \quad \sigma = 1, 2, \dots, s, \quad \begin{aligned}(21·4)\\(21·5)\end{aligned} \end{cases}$$

where $t_1, t_2, \dots, t_n$ are real numbers satisfying

$$\prod_{\rho=1}^{r} t_\rho \prod_{\sigma=1}^{s} (t_{r+\sigma}^2 + t_{r+s+\sigma}^2) = 1. \tag{21·6}$$

Further let $\quad r \geqslant 0, \quad s \geqslant 0, \quad r+s > 1.$

Then K is of rank $r+s-1$ with respect to Γ.

Proof. An arbitrary linear manifold M through O of dimension $r+s-1$ can be defined by $n-(r+s-1) = s+1$ independent homogeneous linear equations

$$a_{h1}x_1 + a_{h2}x_2 + \dots + a_{hn}x_n = 0, \quad \text{where} \quad h = 1, 2, \dots, s+1, \tag{21·7}$$

and where the a's are real numbers. Two cases may now be distinguished:

(*a*) First assume that $r > 0$, and that at least one coefficient

$$a_{hk} \text{ with } \quad 1 \leqslant h \leqslant s+1, \; 1 \leqslant k \leqslant r$$

is different from zero, say the coefficient a_{11}.

Then, on solving the equation,

$$a_{11}x_1+a_{12}x_2+\dots+a_{1n}x_n = 0 \tag{21·8}$$

for x_1,

$$x_1 = b_2x_2+\dots+b_nx_n, \tag{21·9}$$

where $b_2, \dots, b_n$ are real numbers; hence there is a positive constant γ such that

$$|x_1| \leqslant \gamma\{x_2^2+\dots+x_n^2\}^{\frac{1}{2}} \tag{21·10}$$

for all points X of M. Put now $t = \gamma^{1/n}t^*$, and apply the automorphism $X' = \Omega X$ defined by

$$x_1' = t^{-(n-1)}x_1,\ x_2' = tx_2,\ \dots,\ x_n' = tx_n, \tag{21·11}$$

that is

$$x_1 = t^{n-1}x_1',\ x_2 = t^{-1}x_2',\ \dots,\ x_n = t^{-1}x_n'. \tag{21·12}$$

Then

$$F(X) = F(X'), \tag{21·13}$$

and from (21·10) it follows that

$$|t^{n-1}x_1'| \leqslant \gamma\left\{\left(\frac{x_2'}{t}\right)^2+\dots+\left(\frac{x_n'}{t}\right)^2\right\}^{\frac{1}{2}}, \tag{21·14}$$

whence

$$t^nF(X')^n \leqslant \gamma\{x_2'^2+\dots+x_n'^2\}^{\frac{1}{2}}\left|\prod_{\rho=2}^{r} x_\rho' \prod_{\sigma=1}^{s} (x_{r+\sigma}'^2+x_{r+s+\sigma}'^2)\right| \leqslant \gamma\{x_2'^2+\dots+x_n'^2\}^{\frac{1}{2}\{1+(r-1)+2s\}}. \tag{21·15}$$

Hence

$$t^nF(X)^n = t^nF(X')^n \leqslant \gamma|X'|^n, \quad |\Omega X| = |X'| \geqslant \gamma^{-1/n}tF(X) = t^*F(X), \tag{21·16}$$

as asserted.

(*b*) Secondly, let either $r = 0$, or assume that $r > 0$, but that all coefficients

$$a_{hk} \quad \text{with} \quad 1 \leqslant h \leqslant s+1,\ 1 \leqslant k \leqslant r$$

vanish.

Then the equations defining M are of the form

$$a_{h,r+1}x_{r+1}+a_{h,r+2}x_{r+2}+\dots+a_{h,n}x_n = 0, \quad \text{where} \quad h = 1, 2, \dots, s+1. \tag{21·17}$$

Arrange the $2s$ co-ordinates $x_{r+1}, x_{r+2}, \dots, x_n$ as s pairs

$$(x_{r+\sigma}, x_{r+s+\sigma}), \quad \text{where} \quad \sigma = 1, 2, \dots, s. \tag{21·18}$$

Since the $s+1$ equations defining M are independent, and since there are only s such pairs of co-ordinates, it must be possible to express at least one such pair of these co-ordinates in terms of the others. Now assume this is the pair (x_{r+1}, x_{r+s+1}), and that on solving for x_{r+1}, x_{r+s+1}, the following equations are obtained:

$$x_{r+1} = \sum_{\sigma=2}^{s}(b_\sigma x_{r+\sigma}+b_\sigma' x_{r+s+\sigma}), \quad x_{r+s+1} = \sum_{\sigma=2}^{s}(c_\sigma x_{r+\sigma}+c_\sigma' x_{r+s+\sigma}), \tag{21·19}$$

where the coefficients b_σ, b_σ', c_σ, c_σ' are real numbers. Hence there is a positive constant γ such that

$$x_{r+1}^2+x_{r+s+1}^2 \leqslant \gamma\sum_{\sigma=2}^{s}(x_{r+\sigma}^2+x_{r+s+\sigma}^2), \tag{21·20}$$

for all points X of M. Put now $t = \gamma^{1/2s} t^{*n/2s}$, and apply the automorphism $X' = \Omega X$ defined by

$$x'_\rho = x_\rho, \quad \text{where} \quad \rho = 1, 2, \ldots, r, \tag{21·21}$$

$$x'_{r+1} = t^{-(s-1)} x_{r+1}, \quad x'_{r+s+1} = t^{-(s-1)} x_{r+s+1}, \tag{21·22}$$

$$x'_{r+\sigma} = t x_{r+\sigma}, \quad x'_{r+s+\sigma} = t x_{r+s+\sigma}, \quad \text{where} \quad \sigma = 2, 3, \ldots, s, \tag{21·23}$$

or conversely,

$$x_\rho = x'_\rho, \quad \text{where} \quad \rho = 1, 2, \ldots, r \tag{21·24}$$

$$x_{r+1} = t^{s-1} x'_{r+1}, \quad x_{r+s+1} = t^{s-1} x'_{r+s+1}, \tag{21·25}$$

$$x_{r+\sigma} = t^{-1} x'_{r+\sigma}, \quad x_{r+s+\sigma} = t^{-1} x'_{r+s+\sigma}, \quad \text{where} \quad \sigma = 2, 3, \ldots, s. \tag{21·26}$$

Then again

$$F(X) = F(X'), \tag{21·27}$$

and from (21·20)

$$t^{2(s-1)}(x'^2_{r+1} + x'^2_{r+s+1}) \leqslant \gamma t^{-2} \sum_{\sigma=2}^{s} (x'^2_{r+\sigma} + x'^2_{r+s+\sigma}) \leqslant \gamma t^{-2} \,|\, X' \,|^2, \tag{21·28}$$

whence

$$t^{2s} F(X')^n \leqslant \gamma \,|\, X' \,|^2 \left| \prod_{\rho=1}^{r} x'_\rho \prod_{\sigma=2}^{s} (x'^2_{r+\sigma} + x'^2_{r+s+\sigma}) \right| \leqslant \gamma \,|\, X' \,|^{2+r+2(s-1)} = \gamma \,|\, X' \,|^n. \tag{21·29}$$

Hence

$$|\, \Omega X \,| = |\, X' \,| \geqslant \gamma^{-1/n} t^{2s/n} F(X) = t^* F(X), \tag{21·30}$$

as asserted.

Up to now it has only been proved that the rank δ of K with respect to Γ is at least $r+s-1$; one now proves that $\delta < r+s$. This is trivial from definition 10 if $s = 0$. Let therefore $s > 0$. Consider the special $(r+s)$-dimensional linear manifold M_0 defined by the equations

$$x_{r+s+\sigma} = 0, \quad \text{where} \quad \sigma = 1, 2, \ldots, s. \tag{21·31}$$

It suffices to prove that, however Ω is chosen in Γ, there is at least one point X of M_0 such that

$$|\, \Omega X \,| < \sqrt{(n+1)}\, F(X). \tag{21·32}$$

There is no loss of generality in assuming that the point X is such that

$$F(X) = 1; \tag{21·33}$$

hence the point $X = (x_1, \ldots, x_r, x_{r+1}, \ldots, x_{r+s}, 0, \ldots, 0)$ of M_0 satisfies the equation

$$\left| \prod_{\rho=1}^{r} x_\rho \prod_{\sigma=1}^{s} x^2_{r+\sigma} \right| = 1, \tag{21·34}$$

but is otherwise arbitrary.

Let now Ω be any element of Γ, and X the point above of M_0. Then the co-ordinates of $X' = \Omega X$ take the form

$$x'_\rho = t_\rho x_\rho, \quad \text{where} \quad \rho = 1, 2, \ldots, r, \tag{21·35}$$

$$x'_{r+\sigma} = t_{r+\sigma} x_{r+\sigma}, \quad x'_{r+s+\sigma} = t_{r+s+\sigma} x_{r+\sigma}, \quad \text{where} \quad \sigma = 1, 2, \ldots, s, \tag{21·36}$$

and where

$$\prod_{\rho=1}^{r} t_\rho \prod_{\sigma=1}^{s} (t^2_{r+\sigma} + t^2_{r+s+\sigma}) = 1. \tag{21·37}$$

Choose now X in M_0 such that

$$x_\rho = t_\rho^{-1}, \quad \text{where} \quad \rho = 1, 2, \ldots, r, \tag{21·38}$$

$$\left.\begin{aligned} x_{r+\sigma} &= (t^2_{r+\sigma}+t^2_{r+s+\sigma})^{-\frac{1}{2}}, \quad (21·39)\\ x_{r+s+\sigma} &= 0, \quad (21·40) \end{aligned}\right\} \quad \text{where} \quad \sigma = 1, 2, \ldots, s;$$

then evidently $F(X) = 1$, as assumed. This choice of X implies that

$$x'_\rho = 1, \quad \text{where} \quad \rho = 1, 2, \ldots, r, \tag{21·41}$$

$$x'^2_{r+\sigma}+x'^2_{r+s+\sigma} = 1, \quad \text{where} \quad \sigma = 1, 2, \ldots, s, \tag{21·42}$$

and so

$$|X'|^2 = r+s < n+1, \tag{21·43}$$

whence

$$|\Omega X| = |X'| < \sqrt{(n+1)} = \sqrt{(n+1)}\,F(X), \tag{21·44}$$

as asserted. This completes the proof.

THEOREM 25. *Let K: $F(X) \leqslant 1$ be a star body of rank δ with respect to Γ, Λ a critical lattice of K, and ϵ an arbitrary positive number. Then there exist $\delta+1$ independent points $P_1, P_2, \ldots, P_{\delta+1}$ of Λ such that*

$$1 \leqslant F(P_\mu) < 1+\epsilon, \quad \textit{where} \quad \mu = 1, 2, \ldots, \delta+1. \tag{21·45}$$

Proof. Let the assertion be false, i.e. assume that there is a critical lattice Λ_0 of K and a positive number ϵ such that all lattice points P_0 of Λ_0 satisfying

$$1 \leqslant F(P_0) < 1+\epsilon \tag{21·46}$$

lie in a certain δ-dimensional linear manifold M containing O.

From theorem 22, there is a positive number t such that every critical lattice Λ of K contains at least one point P such that

$$1 \leqslant F(P) < 1+\epsilon, \quad |P| \leqslant t. \tag{21·47}$$

Further, by the last definition applied with $t^* = t+1$, there exists an automorphism Ω in Γ such that

$$|\Omega X| \geqslant (t+1)\,F(X) \quad \text{for all points } X \text{ in } M. \tag{21·48}$$

Denote now by

$$P_1,\ P_2,\ P_3,\ \ldots$$

the points of Λ_0 for which

$$1 \leqslant F(P_r) < 1+\epsilon, \quad \text{where} \quad r = 1, 2, 3, \ldots; \tag{21·49}$$

by hypothesis, these points belong to M. Then the only points Q_r of the lattice $\Lambda = \Omega\Lambda_0$ satisfying

$$1 \leqslant F(Q_r) < 1+\epsilon \tag{21·50}$$

are those given by

$$Q_r = \Omega P_r, \quad \text{where} \quad r = 1, 2, 3, \ldots, \tag{21·51}$$

and for these points

$$|Q_r| = |\Omega P_r| \geqslant (t+1)\,F(P_r) \geqslant t+1, \tag{21·52}$$

contrary to the existence result (21·47). Hence the assertion is true.

From theorems 24 and 25, it is deduced that if K is the star body of distance function

$$F(X) = |x_1 x_2 \dots x_n|^{1/n}, \tag{21·53}$$

and Λ is any critical lattice of K, then there exist n independent points $P_1, P_2, \dots, P_n$ of Λ such that

$$1 \leqslant F(P_g) < 1 + \epsilon, \quad \text{where} \quad g = 1, 2, \dots, n, \tag{21·54}$$

however small ϵ is chosen. Hence problem A can be solved in this special case, and the answer is in the affirmative.

I am greatly indebted to Professor Mordell for his help with the manuscript.

REFERENCES

Bachmann, P. 1898 *Quadratische Formen.* **1**, Leipzig: B. G. Teubner.
Davenport, H. 1938 *Proc. Lond. Math. Soc.* **44**, 412–431.
Davenport, H. 1939 *Proc. Lond. Math. Soc.* **45**, 98–125.
Davenport, H. 1944 *J. Lond. Math. Soc.* **19**, 13–18.
Hermite, Ch. 1905 *Oeuvres.* **1**, Paris: Gauthier-Villars.
Hlawka, E. 1943 *Math. Z.* **49**, 285–312.
Koksma, J. F. 1936 *Diophantische Approximationen,* pp. 32, 33. Berlin: J. Springer.
Mahler, K. 1943 *J. Lond. Math. Soc.* **18**, 233–238.
Mahler, K. 1944 *J. Lond. Math. Soc.* **19**, 201–205.
Minkowski, H. 1907 *Diophantische Approximationen.* Leipzig: B. G. Teubner.
Minkowski, H. 1911 *Gesammelte Abhandlungen.* **2**. Leipzig: B. G. Teubner.
Mordell, L. J. 1942 *J. Lond. Math. Soc.* **17**, 107–115.
Mordell, L. J. 1943 *Proc. Lond. Math. Soc.* **48**, 198–228.
Mordell, L. J. 1944 *J. Lond. Math. Soc.* **19**, 6–12.
Mordell, L. J. 1945 *Proc. Lond. Math. Soc.* **48**, 339–390.

The Raman spectrum of rock-salt

By R. S. Krishnan

From the Department of Physics, Indian Institute of Science, Bangalore

(*Communicated by Sir C. V. Raman, F.R.S.—Received* 20 *June* 1945)

[Plate 5]

Using a non-luminescent crystal of rock-salt, a quartz spectrograph with a fine slit, and the 2536·5 A resonance radiations of mercury arc as exciter, the Raman effect in rock-salt has been studied. The spectrum exhibits nine distinct Raman lines with frequency shifts 135, 184, 202, 235, 258, 278, 314, 323 and 350 cm.$^{-1}$. The frequency shifts 235 and 184 cm.$^{-1}$ representing conspicuous lines in the Raman spectrum agree as nearly as could be expected with the position of the two subsidiary infra-red absorption maxima observed by Barnes & Czerny with thin films of rock-salt. The principal infra-red absorption frequency of 163 cm.$^{-1}$ is inactive in the Raman effect, but its octave is represented.

The nature of the Raman spectrum to be expected is deduced on the basis of a theory due to Tamm, as also on the basis of another due to Fermi, the vibration spectrum of the rock-salt lattice being taken to be that worked out by Kellermann on the basis of the Born lattice dynamics. The results are altogether of a different nature from those actually observed experimentally in the present investigation. The conclusion is thus reached that the Born lattice dynamics does not correctly picture the vibration spectrum of the rock-salt lattice. On the other hand the observed facts, both in respect of Raman effect and infra-red absorption, fit into the theoretical picture provided by the dynamics of crystal lattices recently worked out by Sir C. V. Raman.

1. Introduction

Rock-salt is the typical ionic crystal, and specimens of it are readily available. It was naturally, therefore, one of the solids with which the earliest attempts to record a Raman spectrum were made. The failure to obtain any results was explained by Schaefer (1929) as due to the fundamental vibration of the rock-salt lattice being inactive in the Raman effect. A little later, however, the employment of a more powerful technique enabled Rasetti (1931) to achieve better success. The spectrum obtained by him was reproduced, together with a microphotometric record, in a paper written jointly with Fermi (Fermi & Rasetti 1931), the latter author contributing some remarks on the theoretical aspects of the case. The spectrum as seen in the reproductions shows some noteworthy features, but no detailed descriptions or measurements were given in the paper. In view of this lack of detail in the published account of Rasetti's experimental work and the fundamental importance of the case, it appeared desirable that the matter should be thoroughly reinvestigated. This has been done and the results are here reported. The facts disclosed by the present investigation are also fully discussed in the paper in the light of the current theories of crystal dynamics and of the Raman effect in solids.

2. Experimental technique

Following Rasetti, the present author has made use of the 2537 A radiation of the mercury arc for exciting the Raman spectrum. The arrangement, however, differed in some details from Rasetti's. In order to excite the 2537 A radiation with great intensity and without self-reversal, the procedure described by Wood (1912) was followed. A vertical type quartz mercury arc was constructed in the laboratory with mercury cathode and an anode of tungsten. The arc was kept continuously evacuated by an efficient pumping system and was operated on 220 V with sufficient resistance to keep the current well below 2 amp. To secure continuity in the operation of the lamp, a considerable self-inductance was also included in the circuit. The cathode bulb of the arc was kept immersed in running water. The discharge inside the quartz tube was pressed against its wall with the aid of a powerful electromagnet.

The specimen of rock-salt to be investigated was cut in the form of a rectangular block and its faces were well polished. The block was mounted with its longer axis parallel to the arc and facing the most intense portion of the discharge inside it. The light scattered transversely within the crystal was focused on the slit of a Hilger (E 3) quartz spectrograph by means of a condensing lens. The 2537 A radiation in the scattered light was effectively suppressed before its entry into the spectrograph by absorption in a column of mercury vapour. In order to prevent the rock-salt from getting unduly heated, a current of air was directed towards it. Under these conditions the temperature of the crystal was maintained below 40° C. In some of the earlier experiments, a 45° quartz prism was employed to enable the light scattered by the rock-salt along its longer axis to be utilized and focused on the slit of the spectrograph. With this arrangement, however, the strongest Raman line of quartz (466 cm.$^{-1}$) came out feebly in the spectrogram (see figures 1 *c* and *e*, plate 5). In the later experiments, any possible complications of this nature were avoided by dispensing with the quartz prism and utilizing the light scattered by the rock-salt along the shorter axis of the block, though this was a less efficient arrangement. In order to get a satisfactory resolution of the details of the spectrum, the slit of the instrument was kept narrow, the width employed being between 0·02 and 0·03 mm. Ilford special rapid plates were used. Exposures of the order of two to three days were usually given to get a reasonably intense spectrogram.

3. LUMINESCENCE OF ROCK-SALT

In the first instance, several specimens of rock-salt from the Punjab mines were tried out. It was noticed that they acquired a fugitive yellow coloration during the exposure to the intense source of ultra-violet radiation. To start with, the specimens were clear and colourless, but after a few hours of exposure the coloration began to develop, the intensity varying from specimen to specimen. The spectrograms obtained with the crystals exhibiting this effect show two prominent luminescent bands. The first of these extends from 2400 A up to 2600 A (see figure 1 *b*, plate 5). The second band extends from 2800 A to the visible region. Figures 1 *b*, *c*, *d* and *e* represent spectrograms taken with four different specimens of rock-salt in which the coloration for the same duration of exposure was in decreasing order of magnitude. The intensity of the first luminescent band varies also in the same order. It is rather unfortunate that the intense portion of this band falls precisely in the same region where the Raman spectrum of rock-salt is recorded. It will be noticed that in figures 1 *b* and *c* the Raman spectrum is masked by the luminescent band with the result that one can see scarcely any detail in it. In figures 1 *d* and *e*, where the band is comparatively of feebler intensity, one can see the Raman spectrum standing out of the background, but not with all its detail. The last two photographs, namely, figures 1 *f* and *g*, represent printings of different density of the same spectrogram taken with a crystal of rock-salt from Baluchistan. With this specimen the Raman spectrum came out clearly without being masked by luminescence.

4. THE RAMAN SPECTRUM

Using the non-luminescent crystal, three different spectrograms were recorded. A microphotometric record of the best of the three is reproduced together with the spectrum itself as figures 2 *a* and *b*, plate 5. The spectrum of the mercury arc is also included as figure 2*c* for purposes of comparison. The region between the exciting 2536·5 A and the faint mercury line 2561·2 A where the Raman spectrum falls, is fortunately quite clear of any mercury lines except an extremely faint one at 2548·2 A, which can with difficulty be seen in the spectrogram and reveals itself by a kink in the microphotometric record. Adjacent to the exciting 2536·5 A line, there is a band extending towards the longer wave-lengths with a clearly masked maximum of intensity displaced from it by about 60 cm.$^{-1}$. Rasetti thought that this was a Raman band. This view, however, does not appear to be correct, as a band with a maximum of intensity in the same position is also seen in the spectrum of the mercury arc though with less intensity. It appears not improbable that this band is of instrumental origin.

The Raman spectra recorded in the present investigation represent a considerable improvement on the work of Rasetti, a fact which may be ascribed to the use in the present work of the smallest practicable slit-width combined with an exposure sufficiently prolonged to record the fainter features properly. As is evident from figures 2 and 3, plate 5, the spectrum consists of a series of distinct Raman lines, of which at least six can be distinguished without difficulty. Their positions are marked in figure 3, their frequency shifts from the exciting line as measured being respectively 135, 184, 235, 278, 314 and 350 cm.$^{-1}$. The line at 235 cm.$^{-1}$ is both sharp and intense and hence is the most prominent feature of the spectrum. The corresponding anti-Stokes line is also clearly visible in the spectrogram. The lines with frequency shifts 184 and 135 cm.$^{-1}$, though weaker, are also sharp. Besides the six lines listed above, the presence of two other feeble lines with frequency shifts 202 and 258 cm.$^{-1}$ clearly resolved from their near neighbours on either side, has been established by a careful scrutiny of the spectrograms and the microphotometric records. The lines with frequency shifts 278, 314 and 350 cm.$^{-1}$ appear rather broad. The microphotometric records suggest that this is due to the presence in them of imperfectly resolved component lines. Between the lines with frequency shifts 314 and 350 cm.$^{-1}$, an inflexion at about 323 cm.$^{-1}$ is seen in the microphotometric records, indicating the presence of still another Raman line with this frequency shift.

5. THE INFRA-RED ABSORPTION SPECTRUM

In view of this wealth of detail in the Raman spectrum, it is of interest to recall the investigations of Barnes (1932) and Barnes & Czerny (1931) on the behaviour of rock-salt in infra-red absorption. By using thin sputtered films, they succeeded in obtaining more detail in the absorption spectrum and also greater accuracy of measurement. The principal absorption peak for rock-salt was located at $61{\cdot}1 \pm 0{\cdot}3\,\mu$ (163 wave numbers), and besides this, the existence of two other subsidiary absorption maxima was also established. These were located at 40·5 and $51\,\mu$ respectively, the former being the more prominent.

We may here remark that there is no Raman line with a frequency shift 163 cm.$^{-1}$, but this is not surprising, since according to the well-known selection rules, for a crystal with full cubic symmetry (Placzek 1934), the fundamental frequencies active in infra-red absorption would be inactive in light-scattering. We might, however, expect the octave of this frequency as a Raman shift. It is doubtful if either of two Raman lines, having frequency shifts 314 and 350 cm.$^{-1}$ respectively, can be explained in this way, in view of the accuracy with which Barnes & Czerny claim to have determined the infra-red active frequency. As remarked earlier, however, the microphotometric records indicate the presence of a Raman line with a frequency shift of 323 cm.$^{-1}$. This agrees nearly enough with the octave of the principal infra-red active vibration frequency. As the subsidiary infra-red absorption maxima are rather weak and appear on the slope of a steeply falling absorption curve, their frequencies (247 and 196 cm.$^{-1}$), as given by Barnes & Czerny, are probably uncertain to some extent. Indeed, it appears likely that they are the same as the frequency shifts (235 and 184 cm.$^{-1}$ respectively) of the two most conspicuous lines in the Raman spectrum.

6. THE VIBRATIONS OF THE ROCK-SALT LATTICE

We shall now consider the significance of the facts ascertained by the present investigation in relation to the theories of the Raman effect in crystals which have been proposed by Mandelstam, Landsberg & Leontowitsch (1930) and by Tamm (1930). As these authors take the Born lattice dynamics as their starting-point, the task of comparing the consequences of their theories with the experimental facts is made much easier by the recent work of Kellermann (1940, 1941) who has calculated the vibration spectrum of the rock-salt lattice on the same basis. Kellermann analyses the frequency distribution of the lattice vibrations by drawing separate curves for the 'acoustic' and 'optical' waves in the crystal, these being subdivided into the 'longitudinal' and 'transverse' sets of waves. Each of the four distribution curves is a continuous one stretching between wide limits of frequency and exhibiting a single broad hump between these limits. The frequency spectrum obtained by summation of the four separate distributions is likewise a continuous one and stretches from zero frequency upwards to about 320 cm.$^{-1}$, exhibiting three humps in its course between these limits. The limiting frequency for acoustic waves of great length is naturally zero; while for the transverse optical waves, Kellermann finds it to be 151 cm.$^{-1}$, and this he identifies with the frequency of the 'residual rays' of rock-salt. This calculated frequency, it may be remarked, is not very wide of the observed infra-red absorption peak at 163 wave numbers. Kellermann offers no interpretation of the subsidiary infra-red absorption maxima observed by Barnes and Czerny at 247 and 196 cm.$^{-1}$.

7. THEORY OF THE FIRST-ORDER RAMAN EFFECT

Following Born, Tamm (1930) assumes that the vibrations of the crystal lattice in the acoustic as well as in the infra-red regions of frequency may be analysed into sets of plane waves extending through the crystal, each of a definite wave-length. The application of the energy and momentum theorems to a collision between a photon and the vibration quantum leads to the formulae

$$\nu_m = \nu_l \pm \nu_\sigma, \tag{1}$$

$$\vec{K}_m = \vec{K}_l \pm \vec{K}_\sigma, \tag{2}$$

where the ν is a frequency and $\vec{K}$ a reciprocal wave vector; the suffixes l, m and σ refer respectively to the incident light, the scattered light and the vibrational waves. Since the difference between K_m and K_l is never great, it follows from equation (2) that whenever there is an observable scattering, the wave-length of the lattice vibrations giving rise to it is comparable with the wave-length of the incident light and hence large compared with the lattice spacings in the crystal. Hence, the only vibrations which can manifest themselves in the Raman effect are those having such large phase wave-lengths.

The physical meaning of the relations stated in (1) and (2) becomes clearer when we consider the origin of the scattered radiations. Each volume element in the crystal is optically polarized under the influence of the electric field of the light waves traversing it, and this polarization is varied periodically by the vibrations of the lattice. Hence arises an optical stratification of the medium having the same spacing as the phase waves of the mechanical vibrations. According to equation (2), the scattering of light with altered frequency may be interpreted as a kind of reflexion of the incident light waves by these optical stratifications. Alternatively, we may regard each volume element in the crystal pulsating with frequency ν_σ as a source of secondary light waves of altered frequency ν_m as indicated by (1). The optical pulsations in individual volume elements located everywhere in the crystal have coherent phase relationships and therefore also the scattered radiations originating from them. Accordingly, these cancel out by mutual interference except in the special circumstance indicated by equation (2); this equation is the condition for the agreement in phase of the secondary radiations from the different volume elements and their mutual reinforcement in the direction of observation.

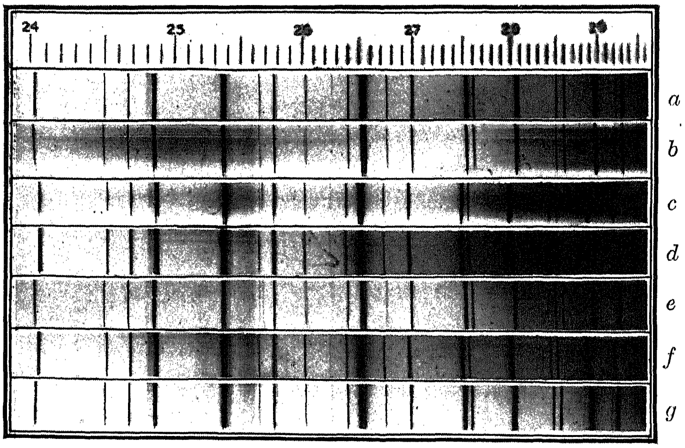


FIGURE 1

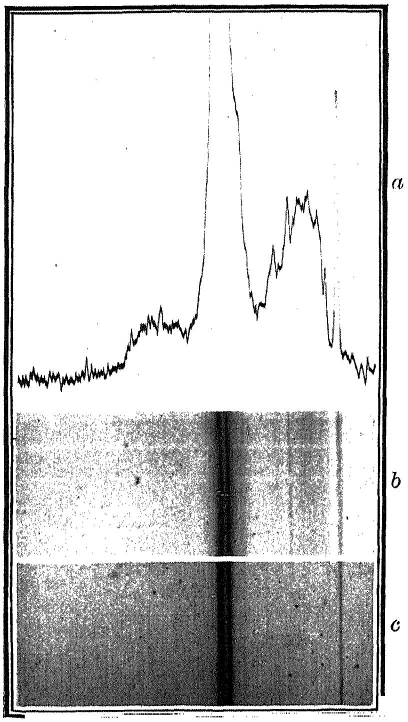


FIGURE 2

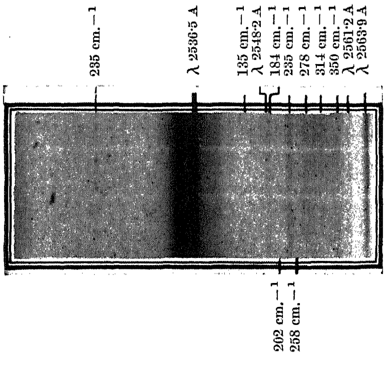


FIGURE 3

8. THEORY OF THE SECOND-ORDER RAMAN EFFECT

Coming now to the case of rock-salt, the changes of frequency due to the reflexion of the light waves by the acoustic waves of appropriate length would be too small to be recorded by a prismatic spectrograph. On the other hand, the limiting frequency for the optical waves corresponds to a vibration active in the infra-red but inactive in light-scattering. Hence, on the basis of the Tamm theory, rock-salt can give no observable Raman effect so long as the oscillations of the lattice and the resulting optical stratifications are assumed to be strictly of a harmonic type. Thus, to find an explanation for the Raman spectrum which is actually recorded with rock-salt, it becomes necessary to consider the second-order effects which arise when the lattice oscillations are anharmonic and/or when the variations of optical polarizability are not exactly proportional to the atomic displacements. The optical stratifications would then not be of the simple harmonic type, either in respect of time variation, or in respect of space distribution. Considering a single set of plane waves of specified frequency and wave-length, a Fourier analysis of the optical stratifications arising from it in respect of time would yield additional terms (a constant term, octave, etc.), while a Fourier analysis in respect of the space distribution would similarly give a constant term, stratifications of half wave-length, etc. When two or more vibrations of different frequencies are superposed, apart from the overtones of the individual vibrations, combinational frequencies would also be present in the variations of optical polarizability. Corresponding complications would also arise in respect of the space variations.

The general features of the Raman effect to be expected when the second-order terms are taken into account can readily be inferred on the basis of the ideas underlying Tamm's treatment of the first-order phenomena. An observable scattering of light with altered frequency can only arise when an optical stratification of corresponding frequency and of appropriate spacing exists, thereby enabling the photon and the crystal to exchange energy and momentum. It is evident that on this basis, the constant terms in the Fourier expansion of the space distribution can give no observable effect and may be ignored. Further, it follows that a second-order Raman spectrum exhibiting overtones and combinations of the lattice frequencies would be observed only in respect of the lattice vibrations which satisfy equations of the same general type as (1) and (2) but suitably modified. We may, for instance, expect the octaves of lattice frequencies to appear as Raman shifts when

$$\nu_m = \nu_l \pm 2\nu_\sigma, \tag{3}$$

$$\vec{K}_m = \vec{K}_l \pm 2\vec{K}_\sigma, \tag{4}$$

the factor 2 in equation (4) indicating that the optical stratifications of double frequency have half the phase wave-length of the mechanical vibrations as their spacing.

The theoretical position may thus be summarized by the statement that the only lattice vibrations which can give rise to observable Raman shifts of frequency, either of the first order (fundamentals) or of the second order (overtones and combinations) are the so-called limiting vibrations which have phase wave-lengths large compared with the lattice spacings of the crystal. Since, according to Kellermann, rock-salt has only one limiting optical frequency, it follows that the second-order Raman spectrum should consist exclusively of a single sharp line with a frequency shift of double that frequency. As we have seen, however, the experimental facts are totally different. We are thus compelled to recognize that the theory of Tamm is incompetent to explain the observed behaviour of rock-salt in the Raman effect.

9. AN ALTERNATIVE THEORY

In his paper written jointly with Rasetti, Fermi (1931) has sought to approach the theory of the Raman effect in crystals from a standpoint which is somewhat different from that of Tamm. He considers the crystal to be made up of numerous volume elements each of which includes a great many lattice cells but is, nevertheless, of dimensions small enough to justify neglect of the variations in phase of the incident and scattered light beams inside it. Fermi then proceeds to consider the phases of the lattice vibrations and the resulting variations of optical polarizability within a volume element. It is obvious that if such variations have a constant sign within a volume element, the scattered radiations from the individual lattice cells contained in it would reinforce each other and give an observable effect. *Per contra*, when the phase varies from cell to cell within the element of volume, the scattered radiations from the individual cells would cancel out each other's effect, and thus give a negligible resultant effect.

The approach suggested by Fermi would obviously be justified only if each volume element could be regarded as functioning independently of the others, in other words, if the vibrations of infra-red frequency in adjacent volume elements have no coherent relations of phase. This, it should be remarked, is a fundamental departure from the basic ideas of the Born lattice dynamics. Nevertheless, it is not without interest to follow up the line of thought suggested by Fermi and to investigate what the Raman spectrum of rock-salt would be if the vibrations of its lattice were assumed to be those described in Kellermann's paper. It is readily shown that the position is the same as on Tamm's theory in respect of the first-order effects. The second-order phenomena, however, would be totally different. We are no longer concerned with the optical stratifications which figure so prominently in the Tamm theory. On the other hand, the space-constant but time-variable effects arising from second-order terms, which were previously insignificant, would now play the leading role. As a consequence, it follows that every mode of vibration of the lattice and not merely the vibration of limiting frequency would contribute its quota to the second-order Raman spectrum exhibiting the octaves and combinations of the lattice frequencies.

According to Kellermann, the frequency of the transverse optical waves ranges from 119 to 194 cm.$^{-1}$. Hence, the octaves would appear in the second-order Raman spectrum as a continuous band covering the range of frequency shifts from 238 to 388 cm.$^{-1}$. The longitudinal optical waves should similarly give a continuous band covering the frequency range from 366 to 640 cm.$^{-1}$. If the acoustical waves could make any sensible contribution to the second-order Raman spectrum, this would similarly be in the nature of a continuous band covering the region of frequency shifts from 0 to 362 cm.$^{-1}$, which would partly overlap and intensify the band due to the transverse optical waves. Combinational frequencies would also manifest themselves, but they need not be discussed in detail, since the result of their superposition would not alter the general nature of the spectrum to be expected, namely a continuous band covering the region of frequency shifts from 238 to 640 cm.$^{-1}$ and possibly extending towards lower frequencies, but not exhibiting any fine structure of the kind actually observed.

10. GENERAL CONCLUSIONS

From the preceding discussion, it is clear that neither on the basis of the Tamm theory nor according to the modified approach suggested by Fermi, is it possible to reconcile the consequences of the Born lattice dynamics with the experimental facts of the Raman effect as observed with rock-salt. The discrepancies observed between fact and theory are of a far-reaching character, and extend to every aspect of the observed spectrum. Hence we are left in no doubt about the conclusion to be drawn, namely, that the vibration spectrum of the rock-salt lattice in the infra-red region of frequency, as deduced from the lattice dynamics of Born, bears no resemblance to the actual vibration spectrum.

The normal modes and frequencies of vibration of the atoms in the space-lattice of a crystal have been investigated in a general manner by Sir C. V. Raman (1943) in a recent memoir. The conclusions reached therein regarding the nature of the vibration spectrum are altogether different from those envisaged in the well-known theories of Debye and of Max Born. It is sufficient here to quote the results of the memoir as applied to the particular case of rock-salt, viz. that the lattice structure of this crystal has nine characteristic modes of vibration in the region of infra-red frequencies. These have been described fully (Chelam 1943) and the nature of the modes is such that none of them can manifest itself as a frequency shift in the Raman spectrum of the first order. All the nine modes, however, are allowed as octaves in the second-order Raman spectrum. Only one of the nine modes is active in infra-red absorption. The eight other modes can only appear as octaves and combinations in the infra-red absorption spectrum.

That the Raman spectrum of rock-salt as actually observed is a second-order effect is clear from the fact that it is only recorded under the most intense excitation. Further, as we have seen, the spectrum exhibits nine distinct Raman lines. The frequency shifts of the two most conspicuous of these agree as closely as could be expected with the frequencies of certain infra-red absorption bands. The latter, it may be inferred from their extreme weakness, are octaves and not actual lattice frequencies. Thus, in all its details, the spectroscopic behaviour of rock-salt fits into the theoretical picture provided by the new approach to crystal dynamics. Further investigations would be required to determine with complete certainty the assignment of the frequencies observed spectroscopically to the particular modes of vibration possible.

In conclusion, the author wishes to express his indebtedness to Professor Sir C. V. Raman, F.R.S., for kind encouragement and useful discussions.

REFERENCES

Barnes 1932 *Z. Phys.* **75**, 723.
Barnes & Czerny 1931 *Z. Phys.* **72**, 447.
Chelam 1943 *Proc. Ind. Acad. Sci.* A, **18**, 257.
Fermi & Rasetti 1931 *Z. Phys.* **71**, 689.
Kellermann 1940 *Phil. Trans.* A, **238**, 513.
Kellermann 1941 *Proc. Roy. Soc.* A, **178**, 17.
Mandelstam, Landsberg & Leontowitsch 1930 *Z. Phys.* **60**, 334.
Placzek 1934 *Marx, Handbuch der Radiologie*, **6**, part 2.
Raman 1943 *Proc. Ind. Acad. Sci.* A, **18**, 237.
Rasetti 1931 *Nature*, **127**, 626.
Schaefer 1929 *Z. Phys.* **54**, 153.
Tamm 1930 *Z. Phys.* **60**, 345.
Wood 1912 *Phil. Mag.* **50**, 761.

DESCRIPTION OF PLATE

FIGURE 1. Spectrograms, two-fold enlarged: *a*, mercury arc and wave-length scale; *b*, *c*, *d*, *e*, scattering and luminescence in four different specimens of rock-salt; *f* and *g*, Raman spectra of non-luminescent crystal.

FIGURE 2. Spectrograms, six-fold enlarged: *b*, Raman spectrum; *a*, its microphotometric record; *c*, mercury arc comparison. *Note.* The diffuse band adjoining the exciting line to its right is seen both in *b* and *c* and is thus not a Raman band.

FIGURE 3. Raman spectrum nine-fold enlarged and with frequency shifts marked.

On Dirac's theory of quantum electrodynamics: the interaction of an electron and a radiation field

BY C. JAYARATNAM ELIEZER

Cosmic Ray Research Unit, Indian Institute of Science, Bangalore; and University of Ceylon

(*Communicated by P. A. M. Dirac, F.R.S.—Received* 27 *June* 1945)

The paper considers the interaction of an electron with a radiation field in the higher approximations. Denoting by ψ_n that part of the wave function which involves the nth power of the charge e and by $\psi_{n,m}$ that part of ψ_n which refers to m photons, it is found that for a general solution $\psi_0, \psi_1, \psi_2, \psi_3$ are all finite and ψ_4 has finite parts $\psi_{4,4}$ and $\psi_{4,2}$ but divergent $\psi_{4,0}$. For the particular solution which corresponds to outgoing waves of the electron, all divergent integrals disappear. $\psi_{n,m}$ is zero for all m different from n, and the value of $\psi_{n,n}$ can be written down easily. The solution makes it now possible to consider exactly processes involving any number of photons, whereas previously exact calculations had to be confined to at most two photons.

INTRODUCTION

It is well known that the occurrence of divergent integrals in the quantum electrodynamics of Heisenberg & Pauli (1929 *a*, *b*) imposes severe restrictions on the range of application of their theory. These divergent integrals appear when we solve the wave equation. Thus, for the motion of an electron in an electromagnetic field, if we try for the wave function a solution of the form

$$\psi = \psi_0 + e\psi_1 + e^2\psi_2 + \dots,$$

a series in ascending powers of the charge e of the electron, then it is found that ψ_0 and ψ_1 are finite, but that ψ_2 contains divergent integrals. In two recent papers, Dirac (1939, 1942) has given a method of eliminating the divergent integrals occurring in ψ_2. Dirac's form of quantum electrodynamics differs from that of Heisenberg and Pauli in two respects. It involves a certain λ limiting process corresponding to an analogous limiting process necessary in classical electrodynamics to express the equations of motion in Hamiltonian form. Further, the representation used is different in that the wave function involves certain field functions as co-ordinates. This new representation requires for its interpretation both positive-energy photons and negative-energy photons, whereas in the representation of Heisenberg and Pauli positive-energy photons alone suffice.

It is the object of this paper to continue with the solution of the wave equation of Dirac's quantum electrodynamics to higher approximations in the perturbation theory, and so examine the higher approximation terms in the interaction of an electron with a radiation field. These solutions are important in that they will enable us to consider many physical processes to a higher degree of approximation than has been possible so far. In particular, the effect of radiation damping on the cross-section for scattering in the Compton effect can be calculated.

THE WAVE EQUATION AND ITS SOLUTION. FIRST AND SECOND ORDER TERMS

Let us consider the motion of an electron in an electromagnetic field. Suppose that initially the electron is at rest. The problem, in which the electron is initially not at rest, can be reduced to this case by means of an appropriate Lorentz transformation. The wave equation is

$$[p_0 - \boldsymbol{\alpha} . (\mathbf{p} - e\mathbf{A}) - m\beta]\,\psi = 0, \tag{1}$$

where m is the mass and e the charge of the electron, the four-vector x gives the time co-ordinate x_0 and the space co-ordinates $\mathbf{x}$ of the electron; p_0, $\mathbf{p}$ are the energy and momentum operators given by

$$p_0 = i\hbar\,\partial/\partial x_0, \quad \mathbf{p} = -i\hbar\,\partial/\partial\mathbf{x};$$

$\boldsymbol{\alpha}$ and β are the usual Dirac matrices. The wave function ψ is taken as a matrix with four rows and four columns, and not as a column matrix, because this representation is more convenient for averaging over the initial states. The units are so chosen that the velocity of light is unity. Using the notation†

$$(a, b) = a^\mu b_\mu = a_0 b_0 - a_r b_r = a_0 b_0 - \mathbf{a} . \mathbf{b}$$

for the scalar product, the four-vector potential $A_\mu(x)$ is given by the Fourier expansion

$$A_u(x) = (\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(2\pi)^{-1}\,\Sigma\int\{\xi_{k\mu}e^{i(k,x)} + \xi^*_{k\mu}e^{-i(k,x)}\}\,k_0^{-1}\,d\mathbf{k} \quad (\mu = 0, 1, 2, 3), \tag{2}$$

where (i) $\xi_{k\mu}$ and $\xi^*_{k\mu}$ are respectively the operators of emission and of absorption of a photon with energy and momentum given by the four-vector $\hbar k$, these operators being defined only for $k^2 = 0$; (ii) Σ means a summation over both values $\pm\sqrt{(k_1^2 + k_2^2 + k_3^2)}$ for k_0; (iii) the potentials satisfy the commutation relations

$$[A_\mu(x), A_\nu(x')] = -\tfrac{1}{2}g_{\mu\nu}\{\Delta(x - x' + \lambda) + \Delta(x - x' - \lambda)\},$$

where Δ is the Jordan-Pauli Δ function. This leads to the following commutation relations for ξ and ξ^*:

$$\left.\begin{array}{l} \xi_{k\mu}\xi_{k'\nu} - \xi_{k'\nu}\xi_{k\mu} = 0, \quad \xi^*_{k\mu}\xi^*_{k'\nu} - \xi^*_{k'\nu}\xi^*_{k\mu} = 0, \\ \xi^*_{k\mu}\xi_{k'\nu} - \xi_{k'\nu}\xi^*_{k\mu} = -\tfrac{1}{2}g_{\mu\nu}(k_0 + k_0')\,\delta(\mathbf{k} - \mathbf{k}')\,e^{-i(k',\lambda)}, \end{array}\right\} \tag{3}$$

where $\delta(\mathbf{k} - \mathbf{k}')$ stands for $\delta(k_1 - k_1')\,\delta(k_2 - k_2')\,\delta(k_3 - k_3')$ and λ is a small time like four-vector so that

$$\lambda^2 = \lambda_0^2 - \boldsymbol{\lambda}^2 > 0. \tag{4}$$

The commutation relations (3) are different from those in the theory which uses only positive energy photons, in which case instead of the factor $\exp\{-i(k,\lambda)\}$ there is a factor $\cos(k,\lambda)$ in the right-hand side of (3). The relations (3) are more symmetrical, but the right-hand side has an imaginary part, showing that these operators are not self-adjoint. This, however, is not a serious limitation since the Hamiltonian is self-adjoint in the limit $\lambda \to 0$. (iv) A_0 is zero here as we are considering a single electron problem.

---

† Summation from 1 to 3 is implied when Latin suffixes are repeated and from 0 to 3 when Greek suffixes are repeated.

The motion of the electron is described by the solution of the wave equation in the limit $\lambda \to 0$.

To solve the equation (1) we try a solution of the form

$$\psi = \psi_0 + e\psi_1 + e^2\psi_2 + \ldots, \tag{5}$$

a series in the ascending powers of e, assuming that e is small. Then the successive coefficients ψ_0, ψ_1, ... are connected by the equation

$$(p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_n = -(\boldsymbol{\alpha}.\mathbf{A})\,\psi_{n-1}. \tag{6}$$

When $n = 0$ this gives $\qquad (p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_0 = 0.$

Hence

$$\psi_0 = u e^{-imx_0/\hbar}, \tag{7}$$

where u is a matrix with four rows and four columns and satisfying the equation

$$(1-\beta)\,u = 0. \tag{8}$$

Denoting $m/\hbar$ by μ, ψ_1 is given by

$$(p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_1 = -(\boldsymbol{\alpha}.\mathbf{A}).ue^{-i\mu x_0}.$$

Substituting for $\mathbf{A}$ from (2), we have

$$(p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_1 = -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(2\pi)^{-1}\,\Sigma\int \alpha_r \xi_{kr}\, e^{i(k,x)-i\mu x_0}\, u k_0^{-1}\, d\mathbf{k}.$$

Hence

$$\begin{aligned}
\psi_1 &= -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(2\pi)^{-1}\,\Sigma\int (m - k_0\hbar + \hbar\boldsymbol{\alpha}.\mathbf{k} - m\beta)^{-1}\alpha_r \xi_{kr}\, e^{i(k,x)-i\mu x_0}\, u k_0^{-1}\, d\mathbf{k} \\
&= -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(2\pi)^{-1}\,\Sigma\int \{(m-k_0\hbar)^2 - \hbar^2k^2 - m^2\}^{-1} \\
&\qquad\qquad \times (m - k_0\hbar - \hbar\boldsymbol{\alpha}.\mathbf{k} + m\beta)\,\alpha_r \xi_{kr}\, e^{i(k,x)-i\mu x_0}\, u k_0^{-1}\, d\mathbf{k}, \\
&= -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(4\pi m)^{-1}\,\Sigma\int (1 + l_s\alpha_s)\,\alpha_r \xi_{kr}\, e^{i(k,x)-i\mu x_0}\, u k_0^{-1}\, d\mathbf{k},
\end{aligned} \tag{9}$$

where (l_1, l_2, l_3) are the direction cosines of the vector $\mathbf{k}$.

ψ_2 is given by $\qquad (p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_2 = -(\boldsymbol{\alpha}.\mathbf{A})\,\psi_1.$

The right-hand side, when simplified by the application of the commutation rules (3), will give two terms—one of second degree and the other of zero degree in the ξ's. We denote by $\psi_{n,m}$ that part of ψ_n which is of the mth degree in the ξ's, that is, the part that refers to m photons. When n is even m is even; and when n is odd m also is odd.

$\psi_{2,0}$ satisfies

$$(p_0-\boldsymbol{\alpha}.\mathbf{p}-m\beta)\,\psi_{2,0} = -(16\pi^2\mu)^{-1}\Sigma\Sigma\iint\alpha_t(1+l_s\alpha_s)\,\alpha_r g_{rt}\delta(\mathbf{k}-\mathbf{k}')$$
$$\times\, e^{i(k-k',x)-i\mu x_0-i(k',\lambda)}\,uk_0^{-1}d\mathbf{k}\,d\mathbf{k}'$$
$$= (16\pi^2\mu)^{-1}\Sigma\int\alpha_r(1+l_s\alpha_s)\,\alpha_r e^{-i\mu x_0-i(k,\lambda)}\,uk_0^{-1}d\mathbf{k}. \qquad (10)$$

In evaluating an integral of the form $\Sigma\int f(k_0,\mathbf{k})\,d\mathbf{k}$ we take

$$\int\{f(k_0,\mathbf{k})+f(-k_0,\,-\mathbf{k})\}\,d\mathbf{k}$$

and integrate over the **k** space by using polar co-ordinates so that

$$d\mathbf{k} = k_0^2dk_0d\Omega = k_0^2\sin\Theta\,dk_0d\Theta\,d\phi.$$

The right-hand side of (10) is then

$$-2i(16\pi^2\mu)^{-1}\int\alpha_r(1+l_s\alpha_s)\,\alpha_r e^{-i\mu x_0}d\Omega\int_{-\infty}^{\infty}k_0\sin k_0(\lambda_0-\mathbf{1}.\boldsymbol{\lambda})\,dk_0,$$

and $\displaystyle\int_{-\infty}^{\infty}k_0\sin k_0(\lambda_0-\mathbf{1}.\boldsymbol{\lambda})\,dk_0 = \left[-\frac{k_0\cos k_0(\lambda_0-\mathbf{1}.\boldsymbol{\lambda})}{\lambda_0-\mathbf{1}.\boldsymbol{\lambda}}+\frac{\sin k_0(\lambda_0-\mathbf{1}.\boldsymbol{\lambda})}{(\lambda_0-\mathbf{1}.\boldsymbol{\lambda})^2}\right]_{-\infty}^{\infty}$,

which can be taken as zero. Hence

$$\psi_{2,0} = 0. \qquad (11)$$

$\psi_{2,2}$ is given by

$$(p_0-\boldsymbol{\alpha}.\mathbf{p}-m\beta)\,\psi_{2,2} = (32\mu\pi^2)^{-1}\Sigma\Sigma\iint\{\alpha_t(1+l_s\alpha_s)\,\alpha_r+\alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\,\xi_{k't}\xi_{kr}$$
$$\times\, e^{i(k+k',x)-i\mu x_0}\,uk_0^{-1}k_0'^{-1}d\mathbf{k}\,d\mathbf{k}', \qquad (12)$$

where the right-hand side has been made symmetrical. Operating on (12) by $p_0+\boldsymbol{\alpha}.\mathbf{p}+m\beta$, we obtain

$$(p_0^2-\mathbf{p}^2-m^2)\,\psi_{2,2} = (32\mu\pi^2)^{-1}\Sigma\Sigma\iint\{m(1+\beta)-\hbar(k_0+k'_0)-\hbar\alpha.(\mathbf{k}+\mathbf{k}')\}$$
$$\times\{\alpha_t(1+l_s\alpha_s)\,\alpha_r+\alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\,\xi_{k't}\xi_{kr}\,e^{i(k+k',x)-i\mu x_0}\,uk_0^{-1}k_0'^{-1}d\mathbf{k}\,d\mathbf{k}'. \qquad (13)$$

In solving an equation of the form

$$(p_0^2-\mathbf{p}^2-m^2)\,\psi = Fe^{i(a_0x_0-\mathbf{a}\,.\,\mathbf{x})/\hbar} \qquad (14)$$

we note that apart from the solution $(a_0^2-\mathbf{a}^2-m^2)^{-1}Fe^{i(a_0x_0-\mathbf{a}\,.\,\mathbf{x})/\hbar}$ there also exist solutions of the form $\delta(a_0^2-\mathbf{a}^2-m^2)\,Fe^{i(a_0x_0-\mathbf{a}\,.\,\mathbf{x})/\hbar}$, because

$$(a_0^2-\mathbf{a}^2-m^2)\,\delta(a_0^2-\mathbf{a}^2-m^2) = 0.$$

Hence the complete solution of (14) is

$$\psi = \{(a_0^2-\mathbf{a}^2-m^2)^{-1}+\phi\delta(a_0^2-\mathbf{a}^2-m^2)\}\,Fe^{i(a_0x_0-\mathbf{a}\,.\,\mathbf{x})/\hbar}, \qquad (15)$$

where ϕ is an arbitrary function of integration, which must be chosen to suit the special physical conditions of each problem under consideration.

Equation (13) is of the form (14) with $a_0 = \hbar(k_0+k_0'-\mu)$, $\mathbf{a} = \hbar(\mathbf{k}+\mathbf{k}')$. Hence

$$\begin{aligned}(a_0^2-\mathbf{a}^2-m^2)\hbar^2 &= -2\mu(k_0+k_0')+2k_0k_0'(1-\mathbf{l}.\mathbf{l}')\\ &= -2\{\mu-k_0(1-\cos\chi)\}\left\{k_0'+\frac{\mu k_0}{\mu-k_0(1-\cos\chi)}\right\},\end{aligned}$$

where χ is the angle between the directions of $\mathbf{l}$ and $\mathbf{l}'$, so that $\mathbf{l}.\mathbf{l}' = \cos\chi$. Thus

$$\phi\delta(a_0^2-\mathbf{a}^2-m^2) = \frac{\phi\hbar^2}{|2\mu-2k_0(1-\cos\chi)|}\,\delta\left\{k_0'+\frac{\mu k_0}{\mu-k_0(1-\cos\chi)}\right\}, \tag{16}$$

the result being expressed in this form because we are going to integrate first with respect to k_0'.

The simplest possible choice of ϕ would be to take it as zero. Then the solution will describe an incoming plane wave and a spherical standing wave. It is found, as will be shown later, that these solutions lead to divergences in the higher approximations.

The next simple choice of ϕ would be to take it equal to a constant. But in the subsequent work, when we evaluate $\psi_{3,1}$, this form of ϕ is found to lead to divergent integrals of form $\int |w|^{-1} e^{-i\lambda_0 w^{-1}}\,dw$, where the range of integration includes $w = 0$. This integral does not tend to a finite value as $\lambda_0 \to 0$. If, however, we take $\phi = \pm C$ according as $\mu-k_0(1-\cos\chi) \gtrless 0$, then the integrals in $\psi_{3,1}$ have the form

$$\int w^{-1} e^{i\lambda_0 w^{-1}}\,dw,$$

which does tend to a finite value as $\lambda_0 \to 0$. This choice has the physical interpretation that $\phi = +C$ for positive energies of the electron and $\phi = -C$ for negative energies of the electron, because the electron's energy $\mu-k_0-k_0'$ has the same sign as $\mu-k_0(1-\cos\chi)$ when k_0, k_0' are such that the δ function in (16) does not vanish, that is, when

$$\mu(k_0+k_0') = kk_0'(1-\cos\chi).$$

When C is taken to be $+i\pi$ it is found that the solution corresponds to outgoing waves of the electron. For this case it will be seen that all the difficulties of divergent integrals disappear.

It is convenient to employ the following notation: If b denotes a four-vector $(b_0, \mathbf{b})$ let

$$\gamma_b \equiv (\mu-b_0)^2-\mu^2-\mathbf{b}^2 \quad\text{and}\quad R_b \equiv \mu(1+\beta)-b_0-\boldsymbol{\alpha}.\mathbf{b}. \tag{17}$$

Further let

$$w \equiv \cos\chi = l.l', \quad w_0 = 1-\mu/\kappa_0, \quad \sigma = \tfrac{1}{2}\mu\lambda_0. \tag{18}$$

Thus the solution of equation (13) is taken to be

$$\begin{aligned}\psi_{2,2} = (32m\pi^2)^{-1}\Sigma\Sigma\iint\{\gamma_{k+k'}^{-1} \pm C\delta(\gamma_{k+k'})\}\,R_{k+k'}\{\alpha_t(1+l_s\alpha_s)\,\alpha_r+\alpha_r(1+l_s'\alpha_s)\,\alpha_t\}\\ \times \xi_{k't}\xi_{kr}\,e^{i(k+k',x)-i\mu x_0}\,uk_0^{-1}k_0'^{-1}\,d\mathbf{k}\,d\mathbf{k}',\end{aligned} \tag{19}$$

where the sign in $\pm$ is determined by

$$\pm\delta(\gamma_{k+k'}) = \frac{1}{2\mu-2k_0(1-\cos\chi)}\,\delta\left\{k_0'+\frac{\mu k_0}{\mu-k_0(1-\cos\chi)}\right\}. \tag{20}$$

THE HIGHER APPROXIMATIONS

ψ_3 consists of two terms: $\psi_{3,3}$ and $\psi_{3,1}$. $\psi_{3,3}$ can be readily evaluated. We obtain

$$\psi_{3,3} \mp -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(64m\hbar\pi^3)^{-1}\,\Sigma\Sigma\Sigma\iiint\{\gamma^{-1}_{k+k'+k''} \pm C\delta(\gamma_{k+k'+k''})\}\,\{\gamma^{-1}_{k+k'} \pm C\delta(\gamma_{k+k'})\}$$
$$\times R_{k+k'+k''}\,\alpha_q R_{k+k'}\{\alpha_t(1+l_s\alpha_s)\,\alpha_r + \alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\,\xi_{k''q}\,\xi_{k't}\,\xi_{kr}$$
$$\times e^{i(k+k'+k'',x)-i\mu x_0}\,u(k_0\,k'_0\,k''_0)^{-1}\,d\mathbf{k}\,d\mathbf{k}'\,d\mathbf{k}''. \tag{21}$$

$\psi_{3,1}$ is given by

$$(p_0-\alpha.p-m\beta)\,\psi_{3,1}$$
$$= (\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(32m\pi^3)^{-1}\,\Sigma\Sigma\iiint\{\gamma^{-1}_{k+k'} \pm C\delta(\gamma_{k+k'})\}\,\alpha_q R_{k+k'}\{\alpha_t(1+l_s\alpha_s)\,\alpha_r$$
$$+\alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\,g_{qt}\,\delta(\mathbf{k}''-\mathbf{k}')\,\xi_{kr}\,e^{i(k+k'-k'',x)-i\mu x_0-i(k'',\lambda)}\,uk_0^{-1}\,k_0'^{-1}\,d\mathbf{k}\,d\mathbf{k}'\,d\mathbf{k}'',$$

where the right-hand side has been reduced by the use of the commutation relations (3). Integrating with respect to $\mathbf{k}''$ we obtain

$$(p_0-\boldsymbol{\alpha}.\mathbf{p}-m\beta)\,\psi_{3,1} = -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(32m\pi^3)^{-1}\,\Sigma\Sigma\iint\{\gamma^{-1}_{k+k'} \pm C\delta(\gamma_{k+k'})\}\,\alpha_t R_{k+k'}\{\alpha_t(1+l_s\alpha_s)\,\alpha_r$$
$$+\alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\,\xi_{kr}\,e^{i(k,x)-i\mu x_0-i(k',\lambda)}\,uk_0^{-1}\,k_0'^{-1}\,d\mathbf{k}\,d\mathbf{k}'.$$

Hence
$$\psi_{3,1} = -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(32m\pi^3)^{-1}\,\Sigma\int(\hbar\gamma_k)^{-1}\,R_k I\,\xi_{kr}\,e^{i(k,x)-i\mu x_0}\,uk_0^{-1}\,d\mathbf{k}, \tag{22}$$
where

$$I \equiv \Sigma\int\{\gamma^{-1}_{k+k'} \pm C\delta(\gamma_{k+k'})\}\,\alpha_t R_{k+k'}\{\alpha_t(1+l_s\alpha_s)\,\alpha_r + \alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\,e^{-ik'_0\lambda_0}\,k_0'^{-1}\,d\mathbf{k}'. \tag{23}$$

We have taken $\boldsymbol{\lambda}$ as zero, which is permissible from the inequalities (4). To evaluate I we note that if $f(k'_0)$ is any polynomial and a is independent of k'_0 then

$$\Sigma\int_{-\infty}^{\infty}\frac{f(k'_0)}{k'_0-a}\,e^{-ik'_0\lambda_0}\,dk'_0 = f(a)\int_{-\infty}^{\infty}\left\{\frac{e^{-ik'_0\lambda_0}}{k'_0-a}-\frac{e^{ik'_0\lambda_0}}{k'_0+a}\right\}dk'_0$$
$$= 2f(a)\left\{a\int_{-\infty}^{\infty}\frac{\cos\kappa'_0\lambda_0}{\kappa_0'^2-a^2}\,dk'_0 - i\int_{-\infty}^{\infty}\frac{k'_0\sin k'_0\lambda_0}{k_0'^2-a^2}\,dk'_0\right\}$$
$$= -2i\pi f(a)\,e^{-ia\lambda_0}.$$

We have used here the result $\displaystyle\int_{-\infty}^{\infty} x^n e^{ix}\,dx = 0,$

for any positive integer n. Again

$$\Sigma\int f(k'_0)\,\delta(k'_0-a)\,e^{-ik'_0\lambda_0}\,dk'_0 = 2f(a)\,e^{-ia\lambda_0}.$$

Hence
$$\Sigma\int\{(\kappa'_0-a)^{-1}+C\delta(\kappa'_0-a)\}f(\kappa'_0)\,e^{-i\kappa'_0\lambda_0}\,d\kappa'_0 = 2(C-i\pi)f(a)\,e^{-ia\lambda_0}. \tag{24}$$

Thus

$$I = (C-i\pi)\int_{\theta'=0}^{\pi}\int_{\phi'=0}^{2\pi}\frac{\mu\alpha_t}{2k_0(w-w_0)^2}\left\{\mu(1+\beta)-k_0(1+l_s\alpha_s)+\frac{\mu}{w-w_0}(1+l'_s\alpha_s)\right\}$$
$$\times\{\alpha_t(1+l_s\alpha_s)\,\alpha_r+\alpha_r(1+l'_s\alpha_s)\,\alpha_t\}\exp\left(\frac{i\mu\lambda_0}{w-w_0}\right)\sin\theta'\,d\theta'\,d\phi'.$$

When integrating with respect to ϕ' it is convenient to take the direction of $\mathbf{l}$ for the polar axis and then to express the result in tensor form. Thus

$$\left.\begin{aligned}\int_0^{2\pi} l'_t\,d\phi' &= 2\pi w l_t,\\ \int_0^{2\pi} l'_s l'_t\,d\phi' &= \pi(1-w^2)\,\delta_{st}+\pi(3w^2-1)\,l_s l_t.\end{aligned}\right\}\tag{25}$$

Substituting and reducing to a simpler form we obtain

$$I = \int_{-1}^{1}\frac{(C-i\pi)\,\pi\mu}{k_0(w-w_0)^2}\Big[4\mu\alpha_r-k_0\{2\alpha_r(1+l_s\alpha_s)+3l_s\,\alpha_s\,\alpha_r-\alpha_r+wl_p\,\alpha_t\,\alpha_p\,\alpha_r\,\alpha_t+wl_s\,l_p\,\alpha_t\,\alpha_s\,\alpha_r\,\alpha_p\,\alpha_t$$
$$+\frac{\mu}{w-w_0}(2\alpha_r+3l_s\alpha_s\alpha_r+wl_p\,\alpha_t\alpha_r\alpha_p\,\alpha_t+2wl_s\alpha_r\alpha_s-w\alpha_r+w^2\alpha_r)\Big]\exp\left(\frac{i\mu\lambda_0}{w-w_0}\right)dw$$
$$=\int_{-1}^{1}\frac{(C-i\pi)\,\pi\mu}{k_0(w-w_0)^2}\left[4\mu-\kappa_0(1+w)\,(1+l_s\alpha_s)+\frac{\mu}{w-w_0}\{2-w+w^2+(3-w)\,l_s\alpha_s\}\right]$$
$$\times\alpha_r\exp\left(\frac{i\mu\lambda_0}{w-w_0}\right)dw$$
$$=(C-i\pi)\,\pi\mu\int_{-1}^{1}\left[\left(\frac{\mu}{k_0}-1-l_s\alpha_s\right)\frac{1}{w-w_0}+\left(\frac{3\mu}{k_0}-\frac{\mu^2}{k_0^2}-1-l_s\alpha_s\right)\frac{2}{(w-w_0)^2}\right.$$
$$\left.+\frac{\mu}{k_0}\left(2-\frac{\mu}{k_0}+\frac{\mu^2}{k_0^2}+\frac{\mu}{k_0}\,l_s\alpha_s+2l_s\alpha_s\right)\right]\alpha_r\exp\left(\frac{i\mu\lambda_0}{w-w_0}\right)dw,\tag{26}$$

where we have omitted certain terms in l_r and $(1-\beta)$ owing to the relations

$$l_r\xi_{kr}=0,\quad (1-\beta)\,u=0.$$

We can perform the integration with respect to w, substitute the value of I in (22) and show that $\psi_{3,1}$ tends to a finite value as λ_0 tends to zero. But for our purposes here, which is to use this value of $\psi_{3,1}$ to evaluate $\psi_{4,2}$ and $\psi_{4,0}$, we should retain λ_0 to the end of the computations. The important result is that $\psi_{3,1}$ is free from divergent integrals.

To proceed to terms of the fourth order, we note that ψ_4 consists of $\psi_{4,4}$, $\psi_{4,2}$ and $\psi_{4,0}$.

$\psi_{4,4}$ can be written down easily, the expression for it being similar to that of $\psi_{3,3}$ in (21).

$\psi_{4,2}$ is given by

$$(p_0-\boldsymbol{\alpha}.\mathbf{p}-m\beta)\,\psi_{4,2}=-(\tfrac{1}{2}\hbar)^{\frac{1}{2}}\,(2\pi)^{-1}\int\alpha_p\{\xi_{k'''p}\,e^{i(k''',x)}\,\psi_{3,1}+\xi^*_{k'''p}\,e^{-i(k''',x)}\,\psi_{3,3}\}\,k_0'''^{-1}d\mathbf{k}'''.$$

The part arising from $\psi_{3,1}$ is free from further integration. The part from $\psi_{3,3}$ can be reduced by applying the commutation relations (3) and integrating with respect to $\mathbf{k}'''$. We then integrate with respect to k_0'' and then when we integrate with respect to the angle variables we obtain integrals of the form

$$I_n = \lim_{\lambda_0 \to 0} P \int_{-1}^{1} \frac{(w-w_0)^{-n}}{\sqrt{\{(w+a)^2+b^2\}}} \exp\left(\frac{i\mu\lambda_0}{w-w_0}\right) dw \quad (n = 0, 1, 2, 3), \tag{27}$$

where P denotes the Cauchy principal value. These integrals are examined in Appendix I and shown to be all convergent and finite as $\lambda_0 \to 0$. Thus the theory does provide a finite solution $\psi_{4,2}$.

$\psi_{4,0}$ is given by

$$(p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_{4,0} = -(\tfrac{1}{2}\hbar)^{\frac{1}{2}}(2\pi)^{-1}\,\Sigma\int \alpha_p\,\xi^*_{k'p}\,e^{-i(k'x)}\,\psi_{3,1}\,k_0'^{-1}\,d\mathbf{k}'$$

$$= -\Sigma\int (2^8 m\mu\pi^4)^{-1}\,\alpha_r\{\mu(1+\beta) - \kappa_0(1+l_s\alpha_s)\}\,Iue^{-i\mu x_0 - ik_0\lambda_0}\,k_0^{-2}\,d\mathbf{k},$$

where we have used the commutation relations (3) and integrated over $\mathbf{k}'$ space. Operating on this equation by $p_0 + \boldsymbol{\alpha}.\mathbf{p} + m\beta$, we have

$$(p_0^2 - \mathbf{p}^2 - m^2)\,\psi_{4,0} = (1+\beta)\,(2^8\mu\pi^4)^{-1}\,\Sigma\int \alpha_r(1+l_s\alpha_s)\,Iue^{-i\mu x_0 - ik_0\lambda_0}\,k_0^{-1}\,d\mathbf{k}. \tag{28}$$

Substituting for I from (26), using the commutation relations satisfied by $\boldsymbol{\alpha}$ and β, and the relation $\beta u = u$, we obtain

$$(p_0^2 - \mathbf{p}^2 - m^2)\,\psi_{4,0} = (2^7\pi^3)^{-1}(C - i\pi)\,3\,\Sigma\iint\left\{\frac{\mu - 2k_0}{w - w_0} - \frac{2}{k_0}\frac{\mu^2 - 3\mu k_0 + 2k_0^2}{(w-w_0)^2} + \frac{\mu}{k_0^2}\frac{\mu^2 + 4k_0^2}{(w-w_0)^3}\right\}$$

$$\times\, u \exp\left[-i\mu x_0 - ik_0\lambda_0 + \frac{i\mu\lambda_0}{w - w_0}\right] dw\,d\Omega\,dk_0.$$

Integrating with respect to Ω, and using the substitutions

$$w_0 = 1 - \mu/k_0, \quad w - w_0 = 2t, \quad k_0 = \tfrac{1}{2}\mu x \quad \text{and} \quad \sigma = \tfrac{1}{2}\mu\lambda_0,$$

we obtain
$$(p_0^2 - \mathbf{p}^2 - m^2)\,\psi_{4,0} = 3(2^6\pi^2)^{-1}(C - i\pi)\,\mu^2 u e^{-i\mu x_0} Q, \tag{29}$$

where
$$Q = \int_{-\infty}^{\infty}\int_{x^{-1}-1}^{x^{-1}}\left\{\frac{1-x}{t} - \frac{2 - 3x + x^2}{xt^2} + \frac{1+x^2}{x^2t^3}\right\}\exp\left[i\sigma\left(\frac{1}{t} - x\right)\right] dt\,dx. \tag{30}$$

This integral is evaluated in Appendix II, and it is found that

$$Q = \left(\frac{2i}{\sigma} - \frac{1}{\sigma^2}\right)\int_{-\infty}^{\infty} dx + \frac{\pi}{\sigma^2}e^{-2i\sigma}\{(1 - 4i\sigma)\,J_0(2\sigma) + 2(i-1)\,J_1(2\sigma)\}, \tag{31}$$

where J_0 and J_1 are Bessel's functions. We thus see that except for $C = i\pi$ we are led to divergent integrals for $\psi_{4,0}$. For $C = i\pi$, equation (24) shows that

$$\psi_{n,m} = 0 \quad \text{for all} \quad m \neq n. \tag{32}$$

The physical interpretation of taking C as $i\pi$ is that we are taking only those solutions which correspond to outgoing waves of the electron. To show this we follow a method given by Dirac (1935) for the treatment of collision problems, where he shows that for outward moving particles, the solution has a factor

$$(W'-W)^{-1}-i\pi\delta(W'-W). \tag{33}$$

In the case considered by Dirac (1935), it applied only to positive energies and we must verify that it also applies to negative energies. For negative W or W', some alterations in the equations of (Dirac 1935, p. 198), will have to be made. Thus on the right-hand side of equation (27) (Dirac 1935), instead of W' we would have $-W' = |W'|$, while in equation (28) there would still be W'. Then the condition (33) above means that the coefficient of $\exp(iP'r/\hbar)$ vanishes and the term with $\exp(-iP'r/\hbar)$ remains, P' being positive. The phase factor thus will be

$$\exp i(-P'r-W't)/\hbar \quad \text{which is} \quad \exp i(-P'r+|W'|t)/\hbar,$$

and hence the solution corresponds to outgoing waves of the electron, for both positive and negative W. It remains to show that the condition (33) leads to the value $i\pi$ for C. Now

$$W' = \mu-k_0-k_0', \quad |W| = \sqrt{\{(\mathbf{k}+\mathbf{k}')^2+\mu^2\}}.$$

The sign of $\mu-k_0(1-\mathbf{1}.\mathbf{1}')$ is opposite to that of W' (or W) for both positive and negative W'. Hence

$$\frac{1}{(\mu-\kappa_0-\kappa_0')^2-\mu^2-(\mathbf{k}+\mathbf{k}')^2}+\frac{i\pi}{2\mu-2k_0(1-\mathbf{1}.\mathbf{1}')}\delta\left\{k_0'+\frac{\mu k_0}{\mu-k_0(1-\mathbf{1}.\mathbf{1}')}\right\}$$
$$=\frac{1}{W'+W}\left\{\frac{1}{W'-W}-i\pi\delta(W'-W)\right\}.$$

For values of C other than $i\pi$ we see that $\psi_{4,0}$ contains divergent integrals. When C is zero the solution corresponds to an incoming plane wave and a spherical standing wave. That this solution which involves only rational functions as integrand should lead to divergent integrals is unexpected in view of a statement made by Dirac, as quoted by Pauli (1943) that 'every (single or multiple) integral with a rational function as integrand is made convergent for large k's by the simultaneous application of the λ limiting process and the negative energy photons'. We now see that this result is not necessarily true for all integrals with rational integrands, for it is not correct to say that every integration will always lead to rational functions. Integrals with respect to k_0 will be of the form

$$\Sigma\int_{-\infty}^{\infty} f(k_0)\,(k_0-a)^{-1}\,e^{-ik_0\lambda_0}\,dk_0,$$

where a is independent of k_0 but is a function of θ, ϕ, $\mathbf{k}'$, $\mathbf{k}''$, ... This integral has the value $-2i\pi f(a)\,e^{-ia\lambda_0}$ which when integrated over the directions of $\mathbf{k}$ may lead to non-rational functions of $\mathbf{k}'$, $\mathbf{k}''$, ... We see that this does happen when we evaluate the integral I in (23).

DISCUSSION

We thus see that in the general solution with the constant C arbitrary, ψ_0, ψ_1, ψ_2, ψ_3 are all finite, and ψ_4 contains finite parts $\psi_{4,4}$ and $\psi_{4,2}$ but divergent $\psi_{4,0}$. The particular solution with C equal to $i\pi$, that is the solution corresponding to outgoing waves of the electron, is free from divergent integrals to any order of approximation in the perturbation theory. $\psi_{n,n}$ is zero for all m different from n. The value of $\psi_{n,n}$ can be written down easily for any n, the expression for it being similar to that of $\psi_{3,3}$ in (21).

The solutions here obtained are important in that they can be used to consider various processes involving any number of photons, whereas previously all accurate determinations had to be restricted to processes involving two photons at most. For example one can now calculate rigorously the probability of one photon splitting into two or more photons in the presence of an electron, or the probability of annihilation of an electron and a positron with the emission of a number of photons. Again, one can now include exactly the effect of radiation damping on the cross-section for scattering in the Compton effect. The principally important term that is likely to lead to alterations in the Klein-Nishina formula for the scattering cross-section is the term $e^4\psi_{4,2}$ and this is found to be zero. Bhabha (1942) has from general considerations given reasons why there should be modifications in the Klein-Nishina formula. It is just possible that the term $e^4\psi_{4,4}$ may give rise to certain contributions towards the scattering cross-section, since two of the four photons, to which this part of the wave function refers, may be such that they have the same frequency, and one is absorbed and the other is emitted in the same direction, thus giving a process which is indistinguishable from one in which two photons take part. In this way it may be possible to reconcile Bhabha's conclusions with the results of this theory. The examination of this question and the calculations of the modification of the Klein-Nishina formula is now in progress and will be dealt with in a further communication.

This work was begun at Cambridge under the supervision of Professor P. A. M. Dirac and was completed under the supervision of Professor H. J. Bhabha at Bangalore. To them both I express my deepest gratitude. I also thank Professor Pauli whose helpful suggestions by correspondence I greatly appreciate and also the University of Cambridge for the award of the Isaac Newton Studentship.

APPENDIX I

We discuss here the convergence of the integrals (27) and examine their value as λ_0 tends to zero.

If $|w_0| > 1$, that is if $k_0 < \frac{1}{2}\mu$, then clearly the integrand is finite everywhere in the range of integration and there is no difficulty about convergence.

If $|w_0| < 1$, that is if $k_0 > \frac{1}{2}\mu$, then the integrand becomes infinite at the point $w = w_0$. The discussion becomes simpler if we transform the integral by putting $w - w_0 = -u^{-1}$ for $w < w_0$, and $w - w_0 = v^{-1}$ for $w > w_0$. Then

$$AI_n = \lim_{\lambda_0 \to 0} \left\{ (-1)^n \int_p^\infty u^{n-1}(u^2 - 2Bu + C)^{-\frac{1}{2}} e^{-iu\lambda_0} du + \int_q^\infty v^{n-1}(v^2 + 2Bv + C)^{-\frac{1}{2}} e^{iv\lambda_0} dv \right\},$$

where A, B and C are certain constants involving a, b and w_0, $p = (1 + w_0)^{-1}$ and $q = (1 - w_0)^{-1}$. p and q are both positive. We have to take the limit λ_0 tends to zero after and not before integration. It is convenient, for the purpose of evaluating this limit, to expand the integrand as a power series thus,

$$(1 - 2Bx + Cx^2)^{-\frac{1}{2}} = \sum_{r=0}^{\infty} a_r x^r, \quad (1 + 2Bx + Cx^2)^{-\frac{1}{2}} = \sum_{r=0}^{\infty} b_r x^r.$$

Then

$$\int_p^\infty u^{n-1}(u^2 - 2Bu + C)^{-\frac{1}{2}} e^{-iu\lambda_0} du = \lambda_0^{1-n} \int_{p\lambda_0}^\infty (t^2 - 2B\lambda_0 + C\lambda_0^2)^{-\frac{1}{2}} t^{\lambda-1} e^{-it} dt$$

$$= \lambda_0^{-1} \int_{p\lambda_0}^\infty (t/\lambda_0)^{n-2} \left\{ \sum_{r=0}^{\infty} a_r (\lambda_0/t)^r \right\} e^{-it} dt.$$

The validity of integrating the function as a series can be justified. For $m > 1$,

$$\int_{p\lambda_0}^\infty t^{-m} e^{-it} dt = \left[(1-m)^{-1} t^{1-m} e^{-it} + \int i(1-m)^{-1} t^{1-m} e^{-it} dt \right]_{p\lambda_0}^\infty.$$

By successive reduction the last term will be $\int_{p\lambda_0}^\infty t^{-1} e^{-it} dt$, which for small λ_0 is of the order of $\log \lambda_0$. Therefore

$$\lambda_0^{m-1} \int_{p\lambda_0}^\infty t^{-m} e^{-it} dt \to (m-1)^{-1} p^{1-m} \quad \text{as} \quad \lambda_0 \to 0.$$

Hence

$$I_0 = -A^{-1} \sum_{r=0}^{\infty} (r+1)^{-1} (a_r p^{-r-1} + b_r q^{-r-1}) = \int_{p^{-1}}^{-q^{-1}} A^{-1}(1 - 2Bx + Cx^2)^{-\frac{1}{2}} dx$$

$$= \int_{-1}^{1} \{(w+a)^2 + b^2\}^{-\frac{1}{2}} dw = \log \left| [(a+1) + \{(a+1)^2 + b^2\}^{\frac{1}{2}}]/[a - 1 + \{(a-1)^2 + b^2\}^{\frac{1}{2}}] \right|, \tag{34}$$

$$AI_1 = -\int_{p\lambda_0}^\infty t^{-1} e^{-it} dt + \int_{q\lambda_0}^\infty t^{-1} e^{-it} dt - \sum_{r=1}^{\infty} r^{-1}(a_r p^{-r} - b_r q^{-r})$$

$$= i\pi + \log(p/q) - \int_p^\infty \{(u^2 - 2Bu + C)^{-\frac{1}{2}} - u^{-1}\} du + \int_q^\infty \{(v^2 + 2Bv + C)^{-\frac{1}{2}} - v^{-1}\} dv$$

$$= i\pi + \log \left| \{p - B + (p^2 - 2Bp + C)^{\frac{1}{2}}\}/\{q + B + (q^2 + 2Bq + C)^{\frac{1}{2}}\} \right|, \tag{35}$$

$$\begin{aligned}AI_2 &= -p-q+a_1\int_{p\lambda_0}^{\infty} t^{-1}e^{-it}\,dt+b_1\int_{q\lambda_0}^{\infty} t^{-1}e^{-it}\,dt+\sum_{r=2}^{\infty}(r-1)^{-1}(a_r p^{1-r}+b_r q^{1-r}),\\ &= i\pi B-p-q+B\log(q/p)+\int_{p}^{\infty}\{u(u^2-2Bu+C)^{-\frac{1}{2}}-1-Bu^{-1}\}\,du\\ &\qquad+\int_{q}^{\infty}\{v(v^2+2Bv+C)^{-\frac{1}{2}}-1+Bv^{-1}\}\,dv\\ &= i\pi B-\{(p^2-2Bp+C)^{\frac{1}{2}}+(q^2+2Bq+C)^{\frac{1}{2}}\}\\ &\qquad-B\log|\{p-B+(p^2-2Bp+C)^{\frac{1}{2}}\}/\{q+B+(q^2+2Bq+C)^{\frac{1}{2}}\}|.\end{aligned} \tag{36}$$

$$\begin{aligned}2AI_3 &= p^2-q^2+2a_1(p+q)+2a_2\{i\pi+\log(q/p)\}+2\sum_{r=2}^{\infty}(r-1)^{-1}\{b_r q^{1-r}-a_r p^{1-r}\}\\ &= i\pi(C+\tfrac{3}{2}B^2)+(p+3B)(p^2-2Bp+C)^{\frac{1}{2}}-(q-3B)(q^2+2Bq+C)^{\frac{1}{2}}\\ &\qquad+(3B^2-C)\log|\{p-B+(p^2-2Bp+C)^{\frac{1}{2}}\}/\{q+B+(q^2+2Bq+C)^{\frac{1}{2}}\}|.\end{aligned} \tag{37}$$

Thus the integrals (27) are all finite and convergent as λ_0 tends to zero.

APPENDIX II

To determine Q of (24) we can take the principal value of the integral, that is, we take the range of integration with respect to x to be $(-H, H)$ and then we make H tend to infinity. The contribution from the second and third terms in the curly bracket in (24) is seen to be, by direct integration with respect to t,

$$\int_{-H}^{H}\left[\left\{\frac{i(1-5x+3x^2-x^3)}{\sigma x(1-x)}+\frac{1+x^2}{\sigma x^2}\right\}\exp\left[i\sigma\left(\frac{x}{1-x}-x\right)\right]+\frac{3i}{\sigma}-\frac{1+x^2}{\sigma^2x^2}\right]dx. \tag{38}$$

We have taken as zero the value of certain terms which oscillate at infinity about the mean value zero.

To evaluate the first term in Q we note that if we directly integrate with respect to t the result will involve the logarithmic integral, and its subsequent integration with respect to x will prove difficult. To avoid this difficulty we invert the order of integration. It is simpler to first transform to a new variable $y = t^{-1}$ and to cut off those points liable to lead to difficulties of divergence. The integral is thus taken as

$$\lim_{\substack{K\to\infty\\ \epsilon\to 0}}\left\{\int_{-H}^{1-\epsilon}\int_{x}^{\frac{x}{1-x}}+\int_{1+\epsilon}^{H}\left(\int_{-K}^{\frac{x}{1-x}}+\int_{x}^{K}\right)\right\}\frac{1-x}{y}\,e^{i\sigma(y-x)}\,dy\,dx. \tag{39}$$

Inverting the order of integration, which process is seen to be justifiable for the area under consideration, and then taking the limit $K\to\infty$ and $\epsilon\to 0$, and omitting certain terms which vanish as $H\to\infty$, we obtain

$$\int_{-H}^{H} f(y)\,y^{-1}e^{i\sigma y}\,dy-\int_{-\infty}^{H} f\left(\frac{y}{y+1}\right)y^{-1}e^{i\sigma y}\,dy,$$

where $$f(x) = \int(1-x)\,e^{-i\sigma x}\,dx = \{i\sigma^{-1}(1-x)-\sigma^{-2}\}\,e^{-i\sigma x}.$$

Thus (39) gives

$$\int_{-H}^{H}\{i\sigma^{-1}(1-y)-\sigma^{-2}\}y^{-1}\,dy-\int_{-\infty}^{H}\{i\sigma^{-1}y^{-1}(y+1)^{-1}-\sigma^{-2}y^{-1}\}\exp\left[i\sigma\left(y-\frac{y}{1+y}\right)\right]dy$$

$$=-i\sigma^{-1}\int_{-H}^{H}dx+\int_{-H}^{\infty}\{i\sigma^{-1}x^{-1}(1-x)^{-1}-\sigma^{-2}x^{-1}\}\exp\left[i\sigma\left(\frac{x}{1-x}-x\right)\right]dx.$$

Hence from (38) and (39)

$$Q=\int_{-\infty}^{\infty}\left(\frac{2i}{\sigma}-\frac{1+x^2}{\sigma^2x^2}\right)dx+\int_{-\infty}^{\infty}\left\{\frac{i(2-5x+3x^2-x^3)}{\sigma x(1-x)}+\frac{1-x+x^2}{\sigma^2x^2}\right\}\exp\left[i\sigma\left(\frac{x}{1-x}-x\right)\right]dx.$$

This integral is convergent at the origin, and by integrating by parts the terms containing x^2 in the denominator we obtain

$$Q=\left(\frac{2i}{\sigma}-\frac{1}{\sigma^2}\right)\int_{-\infty}^{\infty}dx+\int_{-\infty}^{\infty}\left\{\frac{i(2-5x+7x^2-4x^3+x^4)}{\sigma x(1-x)^2}+\frac{x-1}{\sigma^2x}\right\}\exp\left[i\sigma\left(\frac{x}{1-x}-x\right)\right]dx. \tag{40}$$

The second integral, in spite of its complex appearance, can be evaluated easily as follows. We first change the variable from x to t where $x = 1-t$ and then to ϕ, where $t = e^{\phi}$. It is thus found to be

$$2\sigma^{-2}e^{-2i\sigma}\int_0^{\infty}[\{t^2-i\sigma(t^2+t^{-2})\}\cos\sigma(t+t^{-1})+\{1-i\sigma(1+t^2)\}t\sin\sigma(t+t^{-1})](t^2-1)^{-1}dt$$

$$=2\sigma^{-2}e^{-2i\sigma}\int_0^{\infty}\{2\cosh\phi\,\cos(2\sigma\cosh\phi)+(1-4i\sigma\cosh^2\phi)\sin(2\sigma\cosh\phi)\}\,d\phi$$

$$=2\sigma^{-2e-2i\sigma}\left[\frac{dP}{d\sigma}+P-i\sigma\frac{d^2P}{d\sigma^2}\right],$$

where $$P=\int_0^{\infty}\sin(2\sigma\cosh\phi)\,d\phi=\frac{\pi}{2}J_0(2\sigma),$$

J_0 being the Bessel's function of zero order. Hence

$$\frac{dP}{d\sigma}=-\pi J_1(2\sigma),\quad \frac{d^2P}{d\sigma^2}=-2\pi\{J_0(2\sigma)-(2\sigma)^{-1}J_1(2\sigma)\},$$

where J_1 is the Bessel's function of the first order.

We finally obtain

$$Q=\left(\frac{2i}{\sigma}-\frac{1}{\sigma^2}\right)\int_{-\infty}^{\infty}dx+\frac{\pi}{\sigma^2}e^{-2i\sigma}\{(1-4i\sigma)J_0(2\sigma)+2(i-1)J_1(2\sigma)\},$$

which is the result quoted in the text, and which shows that Q contains divergent integrals.

REFERENCES

Bhabha, H. J. 1942 Presidential Address, Agra Session, National Academy of Sciences (India).
Dirac, P. A. M. 1935 *Quantum mechanics*. Oxford University Press.
Dirac, P. A. M. 1939 *Ann. Inst. Poincaré*, **9**, 13.
Dirac, P. A. M. 1942 *Proc. Roy. Soc.* A, **180**, 1–40.
Heisenberg, W. & Pauli, W. 1929*a* *Z. Phys.* **56**, 1–61.
Heisenberg, W. & Pauli, W. 1929*b* *Z. Phys.* **59**, 168–190.
Pauli, W. 1943 *Rev. Mod. Phys.* **15**, 175–207.

The application of quantum electrodynamics to multiple processes

By C. Jayaratnam Eliezer, *University of Ceylon*

(*Communicated by P. A. M. Dirac, F.R.S.—Received* 12 *September* 1945)

The higher approximation terms of the interaction of an electron and a radiation field, which were obtained by the author in the previous paper, are applied here to investigate the probability of multiple scattering processes. It is shown that the probability of certain of these processes could be large. In particular it is shown that the probability of a photon dividing into a number of photons in the presence of an electron is appreciable under certain circumstances. This result shows that quantum electrodynamics does not disallow, as was believed earlier, the existence of photon showers in which the quanta are emitted simultaneously.

INTRODUCTION

1. The well-known difficulty of occurrence of infinite integrals in the higher approximation terms of the interaction of an electron and a radiation field has made it so far not possible to consider *exactly* various multiple scattering processes. Recently, the author (Eliezer 1946)* has investigated the interaction of an electron and a radiation field, on the basis of Dirac's formulation of quantum electrodynamics (Dirac 1939, 1942), and shown that the higher approximations of this interaction are free from the usual infinite integrals. The three features of this theory which bring about the elimination of these divergent integrals are:

(i) the λ-limiting process, which is used to express the equations of motion in Hamiltonian form, and which introduces a convergence factor of the form $\exp\{-i(k,\lambda)\}$ in the commutation relations;

(ii) the use of negative-energy photons as well as positive-energy photons in the process of second quantization;

(iii) the use of only those solutions of the wave equation which correspond to outgoing waves of the electron.

---

* Referred to here as I.

The usual difficulties in quantum electrodynamics being thus removed, it becomes now possible to investigate exactly various multiple processes involving any number of photons, whereas previously all exact calculations had to be restricted to processes in which at most two photons take part. These multiple processes are important for the following reasons. These are typical quantum effects, and their probability cannot be estimated from the classical theory by using the correspondence principle. Also, the existence of *showers* in cosmic radiation makes it desirable to know if in photon showers the quanta are emitted simultaneously or one after another in a short range. By calculating the order of magnitude of the transition probabilities for these multiple processes we shall be able to see whether these probabilities as given by the present theory of quantum electrodynamics are large enough to explain the phenomenon of photon showers.

In this paper we consider scattering processes in which an electron and three photons are involved. The calculations can then be extended in a straightforward way to processes involving a larger number of photons. We follow here a method which was used by Dirac (1930) to calculate the scattering probability in the Compton effect. It will be shown that the theory does not disallow the possibility of photon showers in which the quanta are emitted simultaneously in a single process. This conclusion is contrary to that of Heitler (1936), who has also considered this problem but who arrives at the result that the theory does not explain the production of showers. We believe that this discrepancy is due to the fact that Heitler restricts all the frequency of *all* the emitted photons to be of the same order of magnitude. If, instead, we permit at least one photon to have a low frequency while the frequencies of the rest can be of any order, then it will be seen that the probability of the multiple processes may become considerable. We shall show that the probability of a multiple process varies as k_0^{-1}, where k_0 corresponds to an emitted photon of low frequency. In this way we see that quantum electrodynamics does support a theory of showers which supposes that all the light quanta are emitted simultaneously.

THE WAVE EQUATION AND ITS SOLUTION

2. Following the notation employed in I, the wave equation for the motion of an electron in an electromagnetic field is

$$[p_0 - eA_0 - \boldsymbol{\alpha}.(\mathbf{p} - e\mathbf{A}) - m\beta]\psi = 0, \tag{1}$$

where A_μ is the four-vector potential of the electromagnetic field. For the present problem A_0 may be taken to be zero. It is convenient to take units in which c, the velocity of light, is unity, and $h/2\pi$, where h is Planck's constant, is also unity. ψ is taken as a matrix with four rows and four columns, instead of the usual column matrix, since this representation is more convenient for averaging over the initial states, which averaging becomes necessary later.

The equation (1) is solved by a perturbation method by expressing the wave function as a series in ascending powers of the charge e of the electron, thus

$$\psi = \psi_0 + e\psi_1 + e^2\psi_2 + \dots. \tag{2}$$

It is assumed that e is small enough for such an expansion to be valid. The successive terms $\psi_0, \psi_1, \dots$ are connected by equations of the form

$$(p_0 - \boldsymbol{\alpha}.\mathbf{p} - m\beta)\,\psi_n = -(\boldsymbol{\alpha}.\mathbf{A})\,\psi_{n-1}. \tag{3}$$

The solution has been investigated in detail in I. It has been shown that if we take only those solutions which correspond to outgoing waves of the electron then all the divergent parts, up to any order n in the approximations, become eliminated. We shall therefore assume that the solution is free from divergences, and consider only that part of the wave function which is appropriate for transition processes in which three photons take part.

In general, the wave function ψ_n will consist of a sum of terms of the type $\psi_{n,n}, \psi_{n,n-2}, \dots, \psi_{n,n-2r}, \dots$, where $\psi_{n,m}$ denotes that part of ψ_n which refers to m photons. Hence, the wave function ψ that should be taken to deal with transition processes in which three photons take part is

$$\psi = e^3\psi_{3,3} + e^5\psi_{5,3} + \dots + e^{2n+1}\psi_{2n+1,3} + \dots. \tag{4}$$

It was shown in I that if we take only those solutions which correspond to outgoing waves of the electron, then

$$\psi_{n,m} = 0 \quad \text{for all } m \neq n. \tag{5}$$

Hence, for the transition process under consideration, we take

$$\psi = e^3\psi_{3,3}. \tag{6}$$

To obtain $\psi_{3,3}$ we may take the vector potential $\mathbf{A}$ of the electromagnetic field expressed in its Fourier components, thus,

$$\mathbf{A} = \mathbf{a}e^{i(k,x)}, \tag{7}$$

where the scalar product notation

$$(a,b) = a_\mu b^\mu = a_0 b_0 - \mathbf{a}.\mathbf{b}$$

is employed, and $k_0/2\pi$ gives the frequency and $\mathbf{k}/2\pi$ the momentum and $\mathbf{a}$ the direction of polarization of the photon, and $k^2 = k_0^2 - \mathbf{k}^2 = 0$.

Suppose that initially the electron is at rest. The problem in which the electron is initially not at rest can be considered by an appropriate Lorentz transformation. The initial wave function is then

$$\psi_0 = ue^{-imx_0}, \tag{8}$$

where u is a matrix of four rows and four columns, which satisfies the equation

$$(1-\beta)u = 0. \tag{9}$$

The wave equations that have to be solved are all of the form

$$(p_0-\boldsymbol{\alpha}.\mathbf{p}-m\beta)\,\psi = e^{i(b,x)-imx_0}, \tag{10}$$

where $b\equiv(b_0, \mathbf{b})$ is a four-vector. It is easily seen that the solution of (10) is

$$\begin{aligned}\psi &= (p_0^2-\mathbf{p}^2-m^2)^{-1}\,(p_0+\boldsymbol{\alpha}.\mathbf{p}+m\beta)\,e^{i(b,x)-imx_0}\\ &= \{(m-b_0)^2-\mathbf{b}^2-m^2\}^{-1}\,(m-b_0-\boldsymbol{\alpha}.\mathbf{b}+m\beta)\,e^{i(b,x)-imx_0},\end{aligned} \tag{11}$$

provided $(m-b_0)^2-\mathbf{b}^2-m^2$ is not zero.

It is convenient to employ the following notation:

$$R_b\equiv m-b_0-\boldsymbol{\alpha}.\mathbf{b}+m\beta, \quad \gamma_b\equiv(m-b_0)^2-\mathbf{b}^2-m^2. \tag{12}$$

By solving the equations (3) successively we readily obtain

$$\psi_{1,1} = -\gamma_k^{-1}R_k(\boldsymbol{\alpha}.\mathbf{a})\,ue^{i(k,x)-imx_0}, \tag{13}$$

$$\psi_{2,2} = \gamma_{k+k'}^{-1}\gamma_k^{-1}R_{k+k'}(\boldsymbol{\alpha}.\mathbf{a}')\,R_k(\boldsymbol{\alpha}.\mathbf{a})\,ue^{i(k+k',x)-imx_0}, \tag{14}$$

$$\psi_{3,3} = -\gamma_{k+k'+k''}^{-1}\gamma_{k+k'}^{-1}\gamma_k^{-1}R_{k+k'+k''}(\boldsymbol{\alpha}.\mathbf{a}'')\,R_{k+k'}(\boldsymbol{\alpha}.\mathbf{a}')\,R_k(\boldsymbol{\alpha}.\mathbf{a})\,ue^{i(k+k'+k'',x)-imx_0}. \tag{15}$$

THE TRANSITION PROBABILITY

3. In general $\psi_{3,3}$ is periodic in the time. Exceptions arise, however, when any of the terms of the denominator of (15) vanish, that is, when any of γ_k, $\gamma_{k+k'}$ and $\gamma_{k+k'+k''}$ vanish. When this happens the solution must be modified.

When γ_k is zero, k_0 is zero; and this case is not of interest for our problem here. The vanishing of $\gamma_{k+k'}$ is the condition of conservation of energy and momentum for a transition process in which the two photons k and k' are emitted or absorbed, according to the sign of k_0 and k_0'. This process is also not of interest to us here, since we are concerned with a process in which three photons take part. The vanishing of $\gamma_{k+k'+k''}$ is the condition of conservation of energy and momentum for a process in which three photons k, k', k'' are emitted or absorbed. It is easily seen that $\gamma_{k+k'}$ and $\gamma_{k+k'+k''}$ cannot both vanish simultaneously for non-zero k_0, k_0' and k_0''.

To determine the solution when any of the γ's vanish we see from the work in paper I that the solution of the equation (10) which is valid for a range in which γ_b is zero is

$$\psi = \frac{1}{p_0'+p_0}\left\{\frac{1}{p_0'-p_0}-i\pi\delta(p_0'-p)\right\}R_b\,e^{i(b,x)-imx_0}, \tag{16}$$

where $p_0' = m-b_0$, $p_0^2 = m^2+\mathbf{b}^2$, and the term in $i\pi\delta(p_0'-p_0)$ is introduced to make the solution correspond to outgoing waves of the electron.

Suppose the three photons concerned with in the problem are such that k_0, k_0', k_0'', $\gamma_{k'+k''}$, $\gamma_{k''+k}$ and $\gamma_{k+k'}$ are all non-zero, but $\gamma_{k+k'+k''}$ is zero. In this case we shall see that $\psi_{3,3}$ increases with the time, which shows that transition processes are taking place with the continual appearance of electrons of energy p_0' and momentum $\mathbf{p}'$, where

$$p_0' = m-(k_0+k_0'+k_0''), \quad \mathbf{p}' = -(\mathbf{k}+\mathbf{k}'+\mathbf{k}''). \tag{17}$$

Since $\gamma_{k+k'+k''} = p_0'^2 - \mathbf{p}'^2 - m^2$ the vanishing of $\gamma_{k+k'+k''}$ is equivalent to

$$p_0'^2 - \mathbf{p}'^2 - m^2 = 0, \tag{18}$$

which may also be written as

$$m(k_0 + k_0' + k_0'') = (k_0'k_0'' - \mathbf{k}' . \mathbf{k}'') + (k_0''k_0 - \mathbf{k}''.\mathbf{k}) + (k_0k_0' - \mathbf{k}.\mathbf{k}'). \tag{19}$$

Probabilities with a physical significance are obtained by supposing that the photons are such that (18) does not hold exactly and that a small correction $\delta p_0'$ is made in p_0' for (18) to be satisfied. Then

$$\delta p_0' = -(p_0'^2 - \mathbf{p}'^2 - m^2)/2p_0'. \tag{20}$$

As in Dirac's paper (1930, p. 370) we obtain as a solution which remains finite as $\delta p_0'$ tends to zero and which is appropriate to consider these transition processes,

$$\psi_{3,3} = \gamma_{k+k'}^{-1}\gamma_k^{-1} R_{k+k'+k''}(\boldsymbol{\alpha} . \mathbf{a}'') R_{k+k'}(\boldsymbol{\alpha} . \mathbf{a}') R_k(\boldsymbol{\alpha} . \mathbf{a}) u e^{-i(p',x)} (1 - e^{-i\delta p_0' x_0})/2p_0' \delta p_0'. \tag{21}$$

The term, which gives rise to the transition process under consideration, is obtained by summing (21) for all possible interchanges of k, k' and k''. Let $(\mathbf{l}, \mathbf{l}', \mathbf{l}'')$ represent unit vectors along the direction of motion and $(\mathbf{m}, \mathbf{m}', \mathbf{m}'')$ unit vectors along the direction of polarization of the three photons. Let $(\mathbf{a}, \mathbf{a}', \mathbf{a}'') = (\kappa\mathbf{m}, \kappa'\mathbf{m}', \kappa''\mathbf{m}'')$. We also define the vectors $(\mathbf{n}, \mathbf{n}', \mathbf{n}'')$ where $\mathbf{n} = \mathbf{l} \times \mathbf{m}$, with similar expressions for $\mathbf{n}'$ and $\mathbf{n}''$. The appropriate wave function is then

$$\psi = e^3 \kappa\kappa'\kappa'' R_{k+k'+k''} U u e^{i(p',x)} (1 - e^{-i\delta p_0' x_0})/2p_0' \delta p_0', \tag{22}$$

where

$$U = \Sigma \gamma_{k+k'}^{-1} \gamma_k^{-1} (\boldsymbol{\alpha} . \mathbf{m}'') R_{k+k'} (\boldsymbol{\alpha} . \mathbf{m}') R_k (\boldsymbol{\alpha} . \mathbf{m}), \tag{23}$$

Σ denoting summation of all terms obtained by interchanging k, k' and k''.

The density of electrons represented by (22) is the diagonal sum of the matrix $\psi\phi$, where ϕ is the Hermitian conjugate of ψ. Hence the density of electrons is

$$\tfrac{1}{2} e^6 (p_0^{-1} \kappa\kappa'\kappa'')^2 \Gamma (1 - \cos \delta p_0' x_0)/\delta p_0'^2, \tag{24}$$

where Γ is the diagonal sum of

$$R_{k+k'+k''} U u \bar{u} \bar{U} \bar{R}_{k+k'+k''}. \tag{25}$$

For further computations it is convenient to express the matrix u in terms of $\boldsymbol{\alpha}$ and β. If we take an initial distribution of electrons with no preferential direction of spin, then u must involve β only, and hence we take, consistent with (9) and without loss of generality,

$$u = 1 + \beta. \tag{26}$$

The initial density of electrons is given by the diagonal sum of $\psi_0\phi_0$ which is easily seen to be 8. The probability of transition is given by the ratio of the final and initial densities. We may express the probability in terms of the intensities I, I' and I'' of the three photon beams by substituting $\kappa^2 = (2\pi)^2 I k_0^{-2}$ and similar expressions. We thus obtain for the probability

$$4e^6\pi^6 (p_0' k_0 k_0' k_0'')^{-2} I I' I'' \Gamma (1 - \cos \delta p_0' x_0)/\delta p_0'^2. \tag{27}$$

To obtain a transition probability with a physical meaning we suppose one of the beam of photons, the k''-photon say, to be not sharply monochromatic, but to have intensity $I''_{k''_0}$ per unit frequency range about the appropriate frequency for transitions. The total transition probability is thus

$$4e^6\pi^6(p'_0 k_0 k'_0 k''_0)^{-2}\, II' I''_{k''_0} \Gamma \int \frac{1-\cos \delta p'_0 x_0}{\delta p'^2_0} \frac{dk''_0}{2\pi}. \tag{28}$$

To evaluate this integral we express the variation of k''_0 in terms of the variations of $\delta p'_0$ by using (20). Then the integral in (28) has the value

$$\frac{p'_0 k''_0}{2\pi\gamma_{k+k'}} \int \frac{1-\cos \delta p'_0 x_0}{\delta p'^2_0}\, d\delta p'_0 = \frac{1}{2}\frac{p'_0 k''_0 x_0}{\gamma_{k+k'}},$$

where we have made use of the condition (19).

Substituting for $I''_{k''_0} = k''^3_0/2\pi$ we get as the probability per unit time of a process in which the k''-photon is spontaneously emitted and the k- and k'-photons stimulated

$$e^6\pi^5 k''^2_0\, II'\Gamma/(4p'_0 k^2_0 k'^2_0 \gamma_{k+k'}). \tag{29}$$

The above calculations will also apply to processes in which any of these photons may be absorbed, provided we change the sign of the frequency corresponding to each photon that is absorbed. Thus from (29) we may deduce the probability of annihilation of an electron and a positron by the emission of three photons, or of the splitting of one photon into two photons, or of the combination of two photons into one photon.

For further discussion we need to know the value of Γ. Its evaluation involves rather long calculation and is done in the Appendix. When we substitute the value of Γ and sum the resulting expression for all directions of polarization of the three photons, (29) takes the form

$$\frac{e^6\pi^5 k''^2_0\, II'}{m^2 k^2_0 k'^2_0 \gamma_{k+k'}} \Sigma\left[\frac{\mu_1}{\gamma^2_{k'+k''}} + \frac{\mu_2}{\gamma_{k+k'}\gamma_{k+k''}}\right], \tag{30}$$

where the μ's are homogeneous expressions of the third degree in m, k_0, k'_0 and k''_0.

If we substitute $I = k^3_0/2\pi$, and similarly for I', we obtain as the probability of a triple spontaneous emission

$$\frac{\pi^3}{4}\frac{e^6}{m^2}\frac{k''^2_0 k_0 k'_0}{\gamma_{k+k'}} \Sigma\left[\frac{\mu_1}{\gamma^2_{k+k'}} + \frac{\mu_2}{\gamma_{k+k'}\gamma_{k+k''}}\right], \tag{31}$$

per unit solid angle of direction of emission of the k-photon about **l**, k'-photon about **l**$'$, and k''-photon about **l**$''$, and also per unit frequency ranges of the k- and k'-photons.

By expressing (31) as a probability per unit energy range for the final electron, and then integrating over all directions of emission, we obtain the total probability of the electron making a transition with emission of three photons. By assuming now that all the negative energy states are occupied except one, and that the unoccupied one corresponds to a positron, we obtain the probability of annihilation of an electron and a positron by the emission of three photons.

Suppose we now consider the probability of the double Compton scattering, that is, the probability of one photon, say the k-photon, splitting into two photons k' and k''. We must then change the sign of k_0 in all the above expressions. The probability is then seen to be

$$\frac{\pi^4}{2}\frac{e^6}{m^2}\frac{k_0' k_0''^2 I}{k_0^2 \gamma_{k+k'}}\Sigma\left[\frac{\lambda_1}{\gamma_{k'+k''}^2}+\frac{\lambda_2}{\gamma_{k'-k}\gamma_{k''-k}}\right], \tag{32}$$

per unit solid angle of direction of emission of k'-photon about $\mathbf{1}'$ and k''-photon about $\mathbf{1}''$, and per unit frequency range of the k'-photon. The λ's are obtained by changing the sign of k_0 in the corresponding μ's.

In estimating the order of magnitude of the probability there are two cases of interest to be considered separately. One is when the frequencies of the emitted photons are of the same order of magnitude. This case has been considered by Heitler, who has shown that the probability of the double scattering is smaller than that of the single process by at least 1/137.

The other case of interest to be considered is when there is no restriction imposed on the smallness of the frequencies of the emitted photons. Suppose the emitted photons are such that the frequency of one of them can be as small as we please while the other frequencies can be of any order consistent with the condition of conservation of energy and momentum. If k_0'' is small, then from the conservation of energy and momentum

$$\gamma_{k'+k''-k} = \gamma_{k'-k} - 2k_0''\{m + k_0(1-\mathbf{1}.\mathbf{1}'') - k_0'(1-\mathbf{1}'.\mathbf{1}'')\} = 0,$$

and therefore $\gamma_{k'-k}$ is also small. The expression (32) is then approximately

$$\pi^4 e^6 k_0' k_0''^2 I \lambda_1''/(2m^2\gamma_{k'-k}^3), \tag{33}$$

where

$$\begin{aligned}\lambda_1'' = {} & 32m\{k_0^2 + k_0'^2 - k_0 k_0'(1-\mathbf{1}.\mathbf{1}'^2)\}(1-\mathbf{1}.\mathbf{1}') - 16(k_0'-k_0)\\ & \times\{2m^2(1+\mathbf{1}.\mathbf{1}'^2) - m(k_0'-k_0)(1-\mathbf{1}.\mathbf{1}')^2 + 2(k_0^2+k_0'^2)(1-\mathbf{1}.\mathbf{1}') + k_0 k_0'(1-\mathbf{1}.\mathbf{1}')(1+\mathbf{1}.\mathbf{1}'^2)\}\\ & + 32\{(k_0'-k_0)(1-\mathbf{1}.\mathbf{1}') - m(1+\mathbf{1}.\mathbf{1}'^2)\}\{\mathbf{1}.(\mathbf{k}'-\mathbf{k})\}^2,\end{aligned} \tag{34}$$

where we have omitted certain terms which vanish in the limit k_0'' tends to zero.

Since k_0'' is small, the expression (33) is rather large, since k_0'' occurs as $k_0''^{-1}$. The order of k_0' is obtained from the condition that $\gamma_{k'-k}$ is small. If the incident photon is such that k_0 is much smaller than the rest mass of the electron then k_0' is of the same order as k_0. But if k_0 is much larger than the rest mass of the electron, and if the angle between $\mathbf{k}$ and $\mathbf{k}'$ is large then k_0' is of the same order as the rest mass, and if the angle is small then k_0' is of the same order as k_0.

Hence the probability of a high-energy photon dividing into photons is large, if one of the emitted photons is of low frequency. The other emitted photon is then at least of moderately large frequency.

In the same way if we calculate the probability of a high-energy photon dividing into three or more photons, we see that this probability could be large provided one at least of the emitted photons is of low frequency. The other emitted photons could be of moderately large frequencies. We see thus that the present quantum electrodynamics does give large probability values for certain multiple processes. In particular, we see that the theory does not disallow, as has been believed earlier, the possibility of photon showers in which a large number of photons are emitted simultaneously.

The author wishes to express his sincere thanks to Professor P. A. M. Dirac, who suggested this problem, for his guidance and supervision.

APPENDIX

To evaluate Γ we need to know the value of $Uu\bar{u}\bar{U}$. From (23)

$$\begin{aligned} Uu &= \Sigma\gamma_{k+k'}^{-1}\gamma_k^{-1}(\boldsymbol{\alpha}.\mathbf{m}'')\,R_{k+k'}(\boldsymbol{\alpha}.\mathbf{m}')\,R_k(\boldsymbol{\alpha}.\mathbf{m})\,(1+\beta) \\ &= \Sigma\gamma_{k+k'}^{-1}(-2mk_0)^{-1}\,(\boldsymbol{\alpha}.\mathbf{m}'')\,R_{k+k'}(\boldsymbol{\alpha}.\mathbf{m}')\,\{m(1+\beta)-k_0(1+\boldsymbol{\alpha}.\mathbf{l})\}\,(\boldsymbol{\alpha}.\mathbf{m})\,(1+\beta) \\ &= (2m)^{-1}\,\Sigma\gamma_{k+k'}^{-1}(\boldsymbol{\alpha}\,.\,\mathbf{m}'')\,R_{k+k'}(\boldsymbol{\alpha}\,.\,\mathbf{m}')\,(1+\boldsymbol{\alpha}\,.\,\mathbf{l})\,(\boldsymbol{\alpha}\,.\,\mathbf{m})\,(1+\beta). \end{aligned} \tag{35}$$

For further simplifications it is convenient if we express the matrices $\boldsymbol{\alpha}$ and β in terms of the Dirac matrices $\boldsymbol{\rho}$ and Pauli matrices $\boldsymbol{\sigma}$, where $\boldsymbol{\alpha} = \rho_1\boldsymbol{\sigma}$, $\beta \doteq \rho_3$, and where the $\boldsymbol{\rho}$'s commute with the $\boldsymbol{\sigma}$'s, and both sets satisfy the same commutation relations as the $\boldsymbol{\alpha}$'s. Also, the following notation is helpful:

$$b_1 \equiv 2(\mathbf{m}'\,.\,\mathbf{m}''), \quad b_2 \equiv \mathbf{m}'\,.\,\mathbf{n}''+\mathbf{m}''\,.\,\mathbf{n}', \quad \mathbf{b} \equiv \mathbf{m}'\times\mathbf{n}''+\mathbf{m}''\times\mathbf{n}', \tag{36}$$

with similar expressions for $(b_1', b_2', \mathbf{b}')$ and $(b_1'', b_2'', \mathbf{b}'')$, obtained by cyclic interchange. Then

$$(\boldsymbol{\alpha}\,.\,\mathbf{m}')\,(1+\boldsymbol{\alpha}\,.\,\mathbf{l})\,(\boldsymbol{\alpha}\,.\,\mathbf{m})+(\boldsymbol{\alpha}\,.\,\mathbf{m})\,(1+\boldsymbol{\alpha}\,.\,\mathbf{l}')\,(\boldsymbol{\alpha}\,.\,\mathbf{m}') = b_1''+i\rho_1 b_2''-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}'',$$

where we have made use of the formula $(\boldsymbol{\sigma}\,.\,\mathbf{a})\,(\boldsymbol{\sigma}\,.\,\mathbf{b}) = \mathbf{a}\,.\,\mathbf{b}+i\boldsymbol{\sigma}\,.\,(\mathbf{a}\times\mathbf{b})$. Hence

$$Uu = (2m)^{-1}\,\Sigma\gamma_{k+k'}^{-1}\,\rho_1(\boldsymbol{\sigma}\,.\,\mathbf{m}'')\,R_{k+k'}(b_1''+i\rho_1 b_2''-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}'')\,(1+\rho_3).$$

Therefore

$$Uu\bar{u}\bar{U} = \tfrac{1}{2}m^{-2}\Sigma(P_1\gamma_{k'+k''}^{-2}+P_2\gamma_{k+k'}^{-1}\gamma_{k+k''}^{-1}), \tag{37}$$

where

$$P_1 = \rho_1(\boldsymbol{\sigma}\,.\,\mathbf{m})\,R_{k'+k''}(b_1+i\rho_1 b_2-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b})\,(1+\rho_3)\,(b_1-i\rho_1 b_2-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b})\,R_{k'+k''}(\boldsymbol{\sigma}\,.\,\mathbf{m})\,\rho_1, \tag{38}$$

$$\begin{aligned} P_2 = {} & \rho_1(\boldsymbol{\sigma}\,.\,\mathbf{m}'')\,R_{k+k''}(b_1''+i\rho_1 b_2''-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}'')\,(1+\rho_3)\,(b_1'-i\rho_1 b_2'-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}')\,R_{k+k'}(\boldsymbol{\sigma}\,.\,\mathbf{m}')\,\rho_1 \\ & +\rho_1(\boldsymbol{\sigma}\,.\,\mathbf{m}')\,R_{k+k'}(b_1'+i\rho_1 b_2'-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}')\,(1+\rho_3)\,(b_1''-i\rho_1 b_2''-\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}'')\,R_{k+k''}(\boldsymbol{\sigma}\,.\,\mathbf{m}'')\rho_1, \end{aligned} \tag{39}$$

with similar expressions for P_1', P_2', P_1'' and P_2''.

For our purpose in this paper it is sufficient to know the value of the P_1's. P_2's have the same order of magnitude as the P_1's,

$$\begin{aligned}P_1 &= \rho_1(\boldsymbol{\sigma}\,.\,\mathbf{m})\,R_{k'+k''}\{b_1^2(1+\rho_3)+(b_2^2+\mathbf{b}^2)(1-\rho_3)+2b_1b_2\rho_2-2b_1\rho_1\boldsymbol{\sigma}\,.\,\mathbf{b}\}\,R_{k'+k''}\,\rho_1(\boldsymbol{\sigma}\,.\,\mathbf{m})\\ &= A(1+\rho_3)+B(1-\rho_3)+C\rho_2+\rho_1\boldsymbol{\sigma}\,.\,\mathbf{D},\end{aligned} \tag{40}$$

where

$$\begin{aligned}A &= 2(k_0'^2+k_0''^2)(1-\mathbf{l}'\,.\,\mathbf{l}'')+2k_0'k_0''\{b_1^2\mathbf{l}'\,.\,\mathbf{l}''+b_2^2+\mathbf{b}^2-b_1\mathbf{b}\,.\,(\mathbf{l}'+\mathbf{l}'')\}\\ &= 2\{k_0'^2+k_0''^2+2k_0'k_0''(1-2\mathbf{m}'\,.\,\mathbf{m}''^2)\}(1-\mathbf{l}'\,.\,\mathbf{l}''),\\ B &= 4m^2b_1^2-4mk_0'b_1(b_1-\mathbf{b}\,.\,\mathbf{l}')-4mk_0''b_1(b_1-\mathbf{b}\,.\,\mathbf{l}'')+2(k_0'^2+k_0''^2)(1-\mathbf{l}'\,.\,\mathbf{l}'')\\ &\quad+2k_0'k_0''\{b_1^2+(b_2^2+\mathbf{b}^2)(\mathbf{l}'\,.\,\mathbf{l}'')-b_1\mathbf{b}\,.\,(\mathbf{l}'+\mathbf{l}'')\}\\ &= 16m^2(\mathbf{m}'\,.\,\mathbf{m}'')^2-8m(k_0'+k_0'')(\mathbf{m}'\,.\,\mathbf{m}'')(\mathbf{m}'\,.\,\mathbf{m}''-\mathbf{n}'\,.\,\mathbf{n}'')+2(k_0'^2+k_0''^2)(1-\mathbf{l}'\,.\,\mathbf{l}'')\\ &\quad-4k_0'k_0''(1-\mathbf{l}'\,.\,\mathbf{l}'')\{(\mathbf{m}'\,.\,\mathbf{m}'')(\mathbf{n}'\,.\,\mathbf{n}'')+(\mathbf{m}'\,.\,\mathbf{n}'')(\mathbf{m}''\,.\,\mathbf{n}')\},\\ C &= 4b_1b_2\{m(k_0'+k_0'')-k_0'k_0''(1-\mathbf{l}'\,.\,\mathbf{l}'')\} = -4(\mathbf{m}'\,.\,\mathbf{m}'')(\mathbf{m}'\,.\,\mathbf{n}''+\mathbf{m}''\,.\,\mathbf{n}')\,\gamma_{k'+k''},\\ \mathbf{D} &= 2\{2(\mathbf{l}'\,.\,\mathbf{m})\,\mathbf{m}-\mathbf{l}'\}\{2k_0'^2(1-\mathbf{l}'\,.\,\mathbf{l}'')-2mk_0'b_1^2+k_0'k_0''(b_1^2+b_2^2+\mathbf{b}^2-2b_1\mathbf{b}\,.\,\mathbf{l}'')\}\\ &\quad+2\{2(\mathbf{l}''\,.\,\mathbf{m})\,\mathbf{m}-\mathbf{l}''\}\{2k_0''^2(1-\mathbf{l}'\,.\,\mathbf{l}'')-2mk_0''b_1^2+k_0'k_0''(b_1^2+b_2^2+\mathbf{b}^2-2b_1\mathbf{b}\,.\,\mathbf{l}')\}\\ &\quad-2b_1\{2(\mathbf{b}\,.\,\mathbf{m})\,\mathbf{m}-\mathbf{b}\}\,\gamma_{k'+k''}\\ &= 4\{(k_0'+k_0'')(1-\mathbf{l}'\,.\,\mathbf{l}'')-4m(\mathbf{m}'\,.\,\mathbf{m}'')^2\}\,[k_0'\{2(l'\,.\,m)\,\mathbf{m}-\mathbf{l}'\}+k_0''\{2(\mathbf{l}''\,.\,\mathbf{m})\,\mathbf{m}-\mathbf{l}''\}]\\ &\quad-4(\mathbf{m}'\,.\,\mathbf{m}'')\{2(\mathbf{b}\,.\,\mathbf{m})\,\mathbf{m}-\mathbf{b}\}\,\gamma_{k'+k''}.\end{aligned} \tag{41}$$

The diagonal sum of

$$R_{k+k'+k''}\{A(1+\rho_3)+B(1-\rho_3)+C\rho_2+\rho_1\boldsymbol{\sigma}\,.\,\mathbf{D}\}\,\bar{R}_{k+k'+k''}$$

is the diagonal sum of

$$R_{k+k'+k''}R_{k+k'+k''}\{A(1+\rho_3)+B(1-\rho_3)+C\rho_2+\rho_1\boldsymbol{\sigma}\,.\,\mathbf{D}\},$$

that is, of

$$\begin{gathered}[2(m-k_0-k_0'-k_0'')\{m(1+\rho_3)-\rho_1\boldsymbol{\sigma}\,.\,(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')\}+(k_0+k_0'+k_0'')^2+(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')^2]\\ \times\{A(1+\rho_3)+B(1-\rho_3)+C\rho_2+\rho_1\boldsymbol{\sigma}\,.\,\mathbf{D}\},\end{gathered}$$

which is

$$\begin{aligned}&8(m-k_0-k_0'-k_0'')\{2mA-(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')\,.\,\mathbf{D}\}+4\{(k_0+k_0'+k_0'')^2+(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')^2\}(A+B)\\ &\qquad= 8(m-k_0-k_0'-k_0'')\{2mA-(k_0+k_0'+k_0'')(A+B)-(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')\,.\,\mathbf{D}\},\end{aligned} \tag{42}$$

where we have made use of the condition $\gamma_{k+k'+k''} = 0$. Hence Γ is of the form

$$\Gamma = 4m^{-2}p_0'\,\Sigma[Q_1\gamma_{k'+k''}^{-2}+Q_2\gamma_{k+k'}^{-1}\gamma_{k+k''}^{-1}], \tag{43}$$

where the Q's are homogeneous expressions of the third degree in m, k_0, k_0' and k_0'',

$$Q_1 = 2mA-(k_0+k_0'+k_0'')(A+B)-(\mathbf{k}+\mathbf{k}'+\mathbf{k}'')\,.\,\mathbf{D}. \tag{44}$$

After substituting for A, B and $\mathbf{D}$ we obtain the total transition probability with either state of polarization for each of the three photons, by adding together the expressions obtained by replacing $\mathbf{m}$ by $\mathbf{n}$ and $\mathbf{n}$ by $-\mathbf{m}$, and so on for the other two photons. The following formulae are helpful: If S denotes the summation over all directions of polarization of the k-photon, with similar meaning for S' and S'', then

$$\begin{aligned} S(\mathbf{a}\,.\,\mathbf{m})\,(\mathbf{b}\,.\,\mathbf{m}) &= S(\mathbf{a}\,.\,\mathbf{n})\,(\mathbf{b}\,.\,\mathbf{n}) = \mathbf{a}\,.\,\mathbf{b} - (\mathbf{l}\,.\,\mathbf{a})\,(\mathbf{l}\,.\,\mathbf{b}), \\ S(\mathbf{a}\,.\,\mathbf{m})\,(\mathbf{b}\,.\,\mathbf{n}) &= (\mathbf{l}, \mathbf{a}\times\mathbf{b}). \end{aligned}$$

Hence it is seen that

$$\begin{aligned} \mu_1 = SS'S''Q_1 = {} & 32m\{k_0'^2 + k_0''^2 + k_0' k_0''(1 - \mathbf{l}'\,.\,\mathbf{l}''^2)\}\,(1 - \mathbf{l}'\,.\,\mathbf{l}'') \\ & - 16(k_0 + k_0' + k_0'')\{2m^2(1 + \mathbf{l}'\,.\,\mathbf{l}''^2) - m(k_0' + k_0'')\,(1 - \mathbf{l}'\,.\,\mathbf{l}'')^2 \\ & + 2(k_0'^2 + k_0''^2)\,(1 - \mathbf{l}'\,.\,\mathbf{l}'') - k_0' k_0''(1 - \mathbf{l}'\,.\,\mathbf{l}'')\,(1 + \mathbf{l}'\,.\,\mathbf{l}''^2)\} \\ & + 32\{(k_0' + k_0'')\,(1 - \mathbf{l}'\,.\,\mathbf{l}'') - m(1 + \mathbf{l}'\,.\,\mathbf{l}''^2)\}\{\mathbf{l}\,.\,(\mathbf{k} + \mathbf{k}')\}\{\mathbf{l}\,.\,(\mathbf{k} + \mathbf{k}' + \mathbf{k}'')\} \\ & - 8\gamma_{k'+k''}(1 + \mathbf{l}'\,.\,\mathbf{l}'')\{\mathbf{l}\,.\,(\mathbf{l}' + \mathbf{l}'')\}\{\mathbf{l}\,.\,(\mathbf{k} + \mathbf{k}' + \mathbf{k}'')\}. \end{aligned}$$

By changing k_0 to $-k_0$, we see that

$$\begin{aligned} \lambda_1'' = {} & 32m\{k_0^2 + k_0'^2 - k_0 k_0'(1 - \mathbf{l}\,.\,\mathbf{l}'^2)\}\,(1 - \mathbf{l}\,.\,\mathbf{l}') \\ & - 16(k_0' - k_0)\{2m^2(1 + \mathbf{l}\,.\,\mathbf{l}'^2) - m(k_0' - k_0)\,(1 - \mathbf{l}\,.\,\mathbf{l}')^2 \\ & + 2(k_0^2 + k_0'^2)\,(1 - \mathbf{l}\,.\,\mathbf{l}') + k_0 k_0'\,(1 - \mathbf{l}\,.\,\mathbf{l}')\,(1 + \mathbf{l}\,.\,\mathbf{l}'^2)\} \\ & + 32\{(k_0' - k_0)\,(1 - \mathbf{l}\,.\,\mathbf{l}') - m(1 + \mathbf{l}\,.\,\mathbf{l}'^2)\}\{\mathbf{l}\,.\,(\mathbf{k}' - \mathbf{k})\}^2, \end{aligned}$$

where certain terms which vanish with k_0'' have been omitted. This is the result quoted in the text.

REFERENCES

Dirac, P. A. M. 1930 *Proc. Camb. Phil. Soc.* **26**, 361.
Dirac, P. A. M. 1939 *Ann. Inst. Poincaré*, **9**, 13.
Dirac, P. A. M. 1942 *Proc. Roy. Soc.* A, **180**, 1.
Eliezer, C. J. 1946 *Proc. Roy. Soc.* A, **187**, 197.
Heitler, W. 1936 *Quantum theory of radiation.* Oxford.

Krypton vacuum wave-length measurements

By T. A. Littlefield, *King's College, Newcastle-upon-Tyne*

(*Communicated by W. E. Curtis, F.R.S.—Received* 28 *July* 1945)

The vacuum wave-lengths of thirty-one lines in the arc spectrum of krypton have been compared directly with the red cadmium standard using a reflecting echelon in vacuum. Uncertainties connected with the dispersion of normal air and the specification of the laboratory air used have thereby been eliminated. Estimates of uncertainty suggest that there is justification in expressing the values to eight significant figures.

Introduction

The most precise wave-length measurements require direct comparison with the cadmium red line, the spectroscopic standard of wave-length. It is not always convenient to refer directly to the primary standard, and the Committee of the International Astronomical Union has therefore specified certain secondary standards based upon the work of a number of independent investigators. These wave-lengths are distributed throughout the spectrum and are required to the highest possible precision. The Fabry-Perot interferometer has been most widely used for this work, but it is generally accepted that the uncertainty in vacuum wave-lengths is greater than that associated with corresponding air wave-lengths. Since such determinations are made with air between the Fabry-Perot plates, the additional uncertainty can be attributed partly to difficulties in the specification of laboratory air (temperature, humidity, carbon dioxide concentration, etc.) and to errors in the generally accepted dispersion values for 'normal air'. These are most evident when comparison is extended over a large wave-length interval, such as would prevail if violet and ultraviolet lines were being compared with the cadmium red standard.

The development of the reflecting echelon grating by Williams (1933) provided an alternative method of precision measurement, and since it is quite easy to work with the grating enclosed in an evacuated chamber, direct comparisons of vacuum wave-lengths became possible, thus eliminating the two uncertainties relating to the condition and dispersion values of laboratory air. The method has already been used by Williams & Middleton (1939) to determine vacuum wave-lengths of some 47 lines in the iron spectrum, and by Drinkwater, Richardson & Williams (1940) to determine the wave numbers of the main lines of H_α and D_α and thereby obtain values for the Rydberg constant. Like these, the present work was undertaken to test the suitability of the reflecting echelon for such work and at the same time determine some new values for vacuum wave-lengths in the krypton arc spectrum.

The work of Meggers, Bruin & Humphreys (1929) showed that many of the lines in the krypton arc spectrum might provide satisfactory secondary standards. The first interferometer work is due to Buisson & Fabry (1913) who used a Fabry-Perot interferometer to determine λλ 5570 and 6871 A. Since then several other investigations have been made all using Fabry-Perot interferometers with gaps as large as 10 cm. and for some of which uncertainties of only 0·0001 A are claimed.

APPARATUS

Full details of the general optical arrangements have been given (Williams 1933). Briefly the echelon mounting (carrying slit, lens, and echelon enclosed in its pressure chamber) is capable of being coupled to a spectrograph so that the axes of the two units are at right angles. Light from the echelon slit S_1 (figure 1) is reflected by the prism P and, since S_1 lies in the focal plane of the echelon lens L, a parallel beam of light passes through the quartz window W of the echelon chamber and falls upon the echelon E and the reference mirrors M_1, M_2. A frontal view (figure 1) looking into the quartz window shows the echelon steps and the two reference mirrors M_1, M_2. The reflected beam returns through the echelon lens L, and passing just above the prism P, is focused upon the spectrograph slit S_2. Two images of the slit S_1 are formed by reflexion from the mirrors M_1, M_2 together with a diffraction pattern depending upon the structure of the light used.

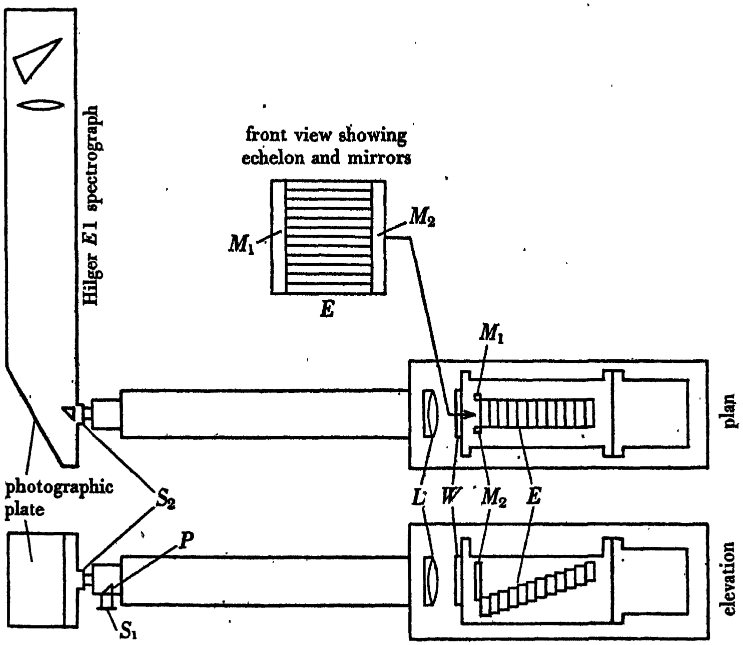


FIGURE 1

The echelon used in the present work differed from that used by Williams (1939, 1940) in that it consisted of 35 instead of 40 steps each approximately 5·999 mm. instead of 6·87 mm. thick. The fringes were focused by a special objective of focal length 164 cm. upon the slit of a Hilger E1 quartz spectrograph. The echelon slit and steps were used in the horizontal position throughout, so that the dispersions of the echelon and spectrograph were crossed. Thus, on being dispersed by the spectrograph, two mirror images and two diffraction fringes corresponding to consecutive orders appear for each wave-length (figure 2). In special cases only one order appears, and F_3 of λ_3 is an example of the single order position. The lengths of the fringes were governed by the width of the spectrograph slit and the choice was dictated by measuring requirements (0·5 mm. was found to be convenient). The choice of echelon slit width, on the other hand, was governed by considerations of lens and echelon resolution and exposure times. The optimum value was determined experimentally by comparing photographs of a water-cooled mercury arc for a series of slit widths—0·025 mm. was found to be a convenient value and has been used in most of the work.

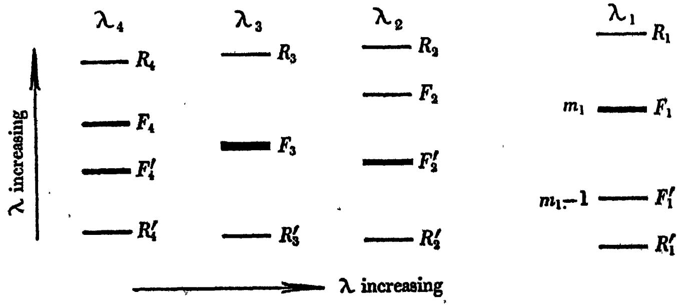


FIGURE 2

With the above arrangement the echelon slit cannot lie on the echelon axis and the echelon lens was therefore adjusted so that its axis bisected the angle between the echelon axis and the echelon slit-lens axis. This adjustment implies that, when the echelon slit is used in either of the other two positions with the dispersion horizontal (i.e. steps vertical), readjustment of the lens axis is desirable.

The Hilger E 1 spectrograph uses a single lens objective and therefore the light falling upon the prism is only strictly parallel for one wave-length. Moreover, the design of the instrument involves working off the axis of the lens and such conditions lead to astigmatism. Normally this does not affect the working of the spectrograph, since it is adjusted and always used with a vertical slit. When, however, it was used in conjunction with the echelon, whose slit was horizontal, it was found necessary to focus the image in a plane 2 mm. in front of the slit for the region $\lambda\lambda$ 8000–4000 A. This difficulty does not arise when the fringes are vertical and the dispersions of the echelon and spectrograph are parallel.

The order separation for the cadmium red standard was determined directly, but, since it was near the single order position, an indirect determination was preferred. The mercury line λ 5461 A was photographed using a water-cooled mercury arc, and as this line has a number of very narrow satellites, the envelope effect is likely to be negligibly small. Moreover, the absence of envelope effect was confirmed by measuring the order separations at a number of different air pressures in the echelon chamber, so that their positions changed from double order to almost single order. The value obtained for the cadmium red standard was 0·1062 cm.

A consideration of temperature and pressure tolerances indicates that these should not vary by more than 0·1° C and 0·1 mm. of mercury. It will be seen later how the procedure adopted provided a further check on these.

The cadmium red line was excited in a hot Schüler tube at 100–125 mA with helium at a pressure of 1 mm. of mercury. Such conditions enabled the red cadmium standard to be photographed in 10–20 min. on an Ilford rapid process panchromatic plate with a slit width of 0·025 mm. An Osram cadmium lamp was also available, but satisfactory exposures required 1½–2 hr. and therefore the Schüler tube was preferred throughout.

Results are based upon photographs from three krypton discharge tubes. One, used in the 'side-on' position, had krypton at a pressure of 1·8 mm. of mercury and was excited by a 3000 V transformer at about 20 mA. The other two tubes were used 'end-on', one small one at 3000 V a.c. and 5 mA, and the other at 2000 V d.c. and 15 mA. Exposures varied from as little as 1 min. with the large tube used 'end-on' to 4 hr. with the first tube used 'side-on' on Ilford rapid process panchromatic plates.

In making a wave-length comparison two photographs of the cadmium standard were taken, followed by a suitable range of krypton exposures, and finally two further exposures with the cadmium standard. During the whole of this period the temperature and pressure were maintained constant within 0·1° C and 0·1 mm. of mercury. Measurement of the cadmium fringe system before and after the krypton exposure provided an optical check on these and upon the stability of the two reference mirrors.

BASIC CONSIDERATIONS

In the present work a vacuum wave-length standard $\lambda_{\text{vac.}} = 6440{\cdot}2491$ A is used which has been obtained from the spectroscopic standard $\lambda_{\text{air}} = 6438{\cdot}4696$ A and the refractive index of 'normal air' 1·000,276,38 at this wave-length, resulting from the work of Barrell & Sears (1939). The dispersion formula given in the latter work for pure air at 15° C and 760 mm. of mercury reduces to

$$(n_{15,760}-1)\,10^6 = 272{\cdot}539+\frac{1{\cdot}5434}{\lambda^2_{\text{vac.}}}+\frac{0{\cdot}0129}{\lambda^4_{\text{vac.}}}.$$

'Normal air', however, contains 0·03 % carbon dioxide and correction for this has been applied along the lines indicated by Pérard (1934) using Perreau's dispersion formula for carbon dioxide:

$$0{\cdot}9997(n_{15,760}-1)\,10^6 = 272{\cdot}457+\frac{1{\cdot}5429}{\lambda^2_{\text{vac.}}}+\frac{0{\cdot}0129}{\lambda^4_{\text{vac.}}} \quad \text{(Barrell \& Sears)},$$

$$0{\cdot}0003(n_{15,760}-1)\,10^6 = 0{\cdot}126+\frac{0{\cdot}0008}{\lambda^2_{\text{air}}} \quad \text{(Perreau)}.$$

Addition gives the combined formula

$$(n_{15,760}-1)\,10^6 = 272{\cdot}583+\frac{1{\cdot}5437}{\lambda^2_{\text{vac.}}}\frac{0{\cdot}0129}{\lambda^4_{\text{vac.}}}.$$

The carbon dioxide correction is sufficiently small to render the use of a combined formula permissible, although it uses vacuum wave-lengths while Perreau's refers to air wave-lengths.

It has already been described how the two fringes and two mirror images appear for each wave-length upon the photographic plate. Before dispersion, the mirror image R on the spectrograph slit is identical for all wave-lengths, and together with the centre of the echelon lens specifies a direction parallel to which light diffracted by the echelon has a constant path difference $2t$. In general the order of interference for R is an integer m plus a fraction f such that

$$(m_1+f_1)\,\lambda_1 = (m_2+f_2)\,\lambda_2 = (m_3+f_3)\,\lambda_3 = 2t.$$

The positions F, F' of fringes on the spectrograph slit correspond to directions for which the orders of interference are successive integers m, $m-1$. The fringes are more widely separated for greater wave-lengths so that on dispersion by the spectrograph $F_1F'_1 > F_2F'_2 > F_4F'_4$, since $\lambda_1 > \lambda_2 > \lambda_4$ (figure 2). The integer m is already known approximately from micrometer measurements of plate thickness t, and the approximate wave-length, while the fractional parts $f_1f_2f_3$, are known accurately from measurements of photographs since $f_1 = R_1F_1/F_1F'_1$, $f_2 = R_2F_2/F_2F'_2$. For wave-lengths (λ_3) near the single order position, in which only one fringe appears, it is first necessary to calculate the order separation by the equation

$$F_3F'_3 = F_1F'_1 . \lambda_3 . R_3R'_3/\lambda_1 . R_1R'_1.$$

In such cases a second mirror image R' is required and, together with R, measures the spectrograph magnification. In the present work this procedure is preferred for all lines irrespective of whether or not they are near the single order position. The order separation for the primary standard is used since it is known with least uncertainty. The exact orders of interference for λ_2, λ_3 are calculated for a range of integers about m_1, and a value for m_1 is finally chosen which for λ_2, λ_3 gives fractional parts most nearly equal to f_2, f_3 respectively. Thus $2t$ is known accurately and other wave-lengths λ_4 are calculated from $\lambda_4 = 2t/(m_4+f_4)$ provided λ_4 is already known with sufficient accuracy to enable m_4 to be calculated without ambiguity.

The above method of calculating results assumes that the dispersion of the reflecting echelon is linear, but this is only strictly true when consideration is restricted to first order terms. Investigation shows that when second order terms are taken into account, the departure from linearity is unlikely to affect results by more than 0·0001 A for the echelon used in the present work.

Williams & Middleton (1939) have shown that the resolving power of a reflecting echelon expressed in wave-numbers is independent of wave-length and equal to $1/2Nt$ cm.$^{-1}$, where t is the thickness of the plate in cm. and N is the number of plates. For the present instrument this gives 0·0238 cm.$^{-1}$ representing 0·0060 A at λ 5000 A.

RESULTS

Results given in table 1 are classified in three groups A, B and C. A are based upon six independent comparisons using each of the three available krypton tubes; B are obtained from three comparisons using a single tube; while C are taken from only two comparisons and tend to be under-exposed. Columns 2 and 3 give vacuum wave-lengths and wave-numbers referred to a cadmium vacuum standard 6440·2491 A without reference to the dispersion formula for air. In column 4 the corresponding wave-lengths in 'normal air' have been calculated by a dispersion formula obtained by combining the Barrell and Sears formula for pure air with Perreau's formula for carbon dioxide. The wave-lengths recommended by the International Astronomical Union are recorded in column 5 and differences are given in the next column. Where I.A.U. values are not available, Jackson's (1936) have been used and are marked (J) accordingly.

The values obtained are generally higher than those recommended by the I.A.U., particularly in the violet, the mean difference for all the wave-lengths being 0·0014 A.

TABLE 1

1 classification	2 vacuum wave-length	3 wave number	4 calculated normal air	5 decimal parts of I.A.U. values	6 *L*-I.A.U.
A	4275·1731	23390·865	4273·9718	·9700	+0·0018
B	4284·1733	23341·726	4282·9696	·9683	+0·0013
B	4287·6930	23322·565	4286·4884	·4873	+0·0011
C	4301·6959	23246·645	4300·4876	·4877	−0·0001
A	4319·7674	23149·394	4318·5543	·5525	+0·0018
A	4320·7953	23143·887	4319·5819	·5797	+0·0022
C	4352·5855	22974·850	4351·3637	·3607	+0·0030
A	4363·8683	22915·448	4362·6435	·6423	+0·0012
A	4377·3524	22844·859	4376·1241	·1220	+0·0021
B	4401·2034	22721·058	4399·9688	·9670	+0·0018
C	4426·4331	22591·553	4425·1918	·1906 (J)	+0·0012
A	4455·1686	22445·839	4453·9197	·9179	+0·0018
A	4464·9435	22396·699	4463·6920	·6902	+0·0018
A	4503·6180	22204·369	4502·3563	·3547	+0·0016
B	5563·7726	17973·416	5562·2286	·2257	+0·0029
A	5571·8371	17947·401	5570·2909	·2895	+0·0014
B	5581·9385	17914·923	5580·3896	·388 (J)	—
B	5651·1308	17695·573	5649·5634	·5628	+0·0006
C	5674·0255	17624·171	5672·4520	·4519 (J)	+0·0001
C	5709·0963	17515·907	5707·5134	·512 (J)	—
B	5834·4770	17139·496	5832·8605	·859 (J)	—
B	5868·3775	17040·485	5866·7519	—	—
A	5872·5441	17028·394	5870·9174	·9158	+0·0016
B	5881·5300	17002·378	5879·9009	·9000 (J)	+0·0009
B	5995·5115	16679·144	5993·8518	·8503	+0·0015
B	6013·8221	16628·360	6012·1575	·1570 (J)	+0·0005
B	6057·8043	16507·631	6056·1279	·1280 (J)	−0·0001
C	6084·5470	16435·077	6082·8635	—	—
C	6238·0772	16030·580	6236·3524	·354 (J)	—
C	6422·8033	15569·525	6421·0289	·029	—
C	6458·0753	15484·489	6456·2914	·291	—

This difference, however, almost disappears when the dispersion formulae due to Kösters & Lampe (1934) or Pérard (1934) are employed, the mean difference being only 0·00016 A in the former case. Presumably one of these formulae was used in calculating the I.A.U. wave-lengths.

An assessment of the accuracy of the work has been attempted by three methods, but only one of these is applicable to B and C class values. It will be appreciated, however, that the performance of the reflecting echelon should be judged on the basis of A class results, for which all three methods are available.

Each plate, having a separate cadmium standard, can be regarded as an independent comparison, and the standard deviation of all lines from their respective means provides an indication of the uncertainty. For ten lines, each appearing on six different plates, the standard deviation is ± 0·0005 A.

Of the lines measured, seven terminate at the IS_4 level and eight at the IS_5 level, and six estimates of the separation of the IS_4 and IS_5 levels are possible. Assuming rigid accuracy of the Ritz combination principle, this provides a useful check on the uncertainty of the wave-lengths involved. The mean inter-level difference is 945·025 cm.$^{-1}$ and the standard deviation of the mean is rather less than ± 0·002 cm.$^{-1}$, which at λ 5000 A corresponds to ± 0·0005 A. Moreover, the above estimate is based upon four B and one C class results; even better agreement is obtained when only A class results are used. The two differences, based entirely upon A class results, differ by 0·001 cm.$^{-1}$ corresponding to only 0·0002 A for the particular wave-lengths concerned ($\lambda\lambda$ 4275–4503 A).

TABLE 2

(4319–4503) = 23149·394 – 22204·369 = 945·025
(4284–4464) = 23341·726 – 22396·699 = 945·027
(4275–4455) = 23390·865 – 22445·839 = 945·026
(5571–5881) = 17947·401 – 17002·378 = 945·023
(5563–5872) = 17973·416 – 17028·394 = 945·022
(5674–5995) = 17624·171 – 16679·144 = 945·027

The third estimate was obtained by making all measurements in duplicate on separate occasions and the mean difference between the settings is 0·0002 cm. For the primary standard, measurements are made on at least two sets of fringes, so that this uncertainty is taken as ± 0·00005 cm. Analysis using the above results, and based on the constants and dimensions of the echelon used, lead to a final uncertainty in a single wave-length comparison of ± 0·0005 A.

Estimates of uncertainty by the above three methods are thus in close agreement and the standard deviation may be set at ± 0·0005 A. As A class results are the mean of six comparisons the standard error is only ± 0·0002 A, which implies that, if a number of further comparisons were made, 95 % of the values obtained would fall in the range ± 0·0004 A about the mean values given. For B and C class results the standard error is placed at ± 0·0003 and ± 0·0005 A respectively, but for these only the first method of assessment is available.

HYPERFINE STRUCTURE

Although krypton is a mixture of three isotopes, only one of these, with a relative abundance of 11·79 %, has odd mass number. As no splitting occurs due to nuclear spin with isotopes of even mass number, the total intensity of the components cannot exceed 12 % of the total intensity of the line. Of the krypton lines examined only λ 5571 A showed structure and this appeared as faint satellites on either side of the main component. As considerable over-exposure of the main component is necessary before the satellites appear, structure was only observed in the three photographs with the large tube used 'end-on' and comparable exposures with the other tubes would have been inconveniently long. In measuring it was found most convenient to obtain the wave-length of each satellite by direct comparison with the cadmium standard, while the wave-length of the main component was determined from a more suitable exposure on the same plate. The structure in wave numbers is +0·104, 0, −0·109 for which the standard deviation is ±0·002 cm.$^{-1}$. It will be seen in table 3 that these values compare favourably with those of other workers, notably Jackson (1936), Kopfermann & Wieth Knudsen (1933) and Humphreys (1931).

TABLE 3

1 Jackson	2 Kopfermann & Wieth Knudsen	3 Humphreys	4 Littlefield
+0·11	+0·111	+0·1073	+0·104
−0·10	−0·120	—	−0·109

DISCUSSION

The present work indicates that a standard deviation of ±0·0005 A is attainable with a reflecting echelon used in the manner recommended by Williams. Thus twenty-six independent comparisons would enable the standard error to be set at ±0·0001 A, the implication of which has already been described. This is perhaps not as good as that usually attributed to the most reliable work with a Fabry-Perot interferometer, although, of course, uncertainties connected with the dispersion of air are eliminated. It is therefore desirable to examine the possibility of increasing the reliability of comparisons with a reflecting echelon, and the experience gained suggests the lines along which such improvement might reasonably be achieved.

Since temperature and pressure affect successive orders equally, the order separation is substantially independent of temperature and pressure variations. It is, therefore, suggested that the order separation for the cadmium red standard might well be regarded as an instrumental constant from which all other order separations should be calculated as required—this procedure has been preferred in the present work.

Calculation of order separation involves a knowledge of the spectrograph magnification and at present the two mirror images provide a convenient means of obtaining it. An uncertainty of ± 0·0002 cm. represents 0·1 % of the mirror image separation (~ 0·2 cm.). By making a separate magnification estimate (which need only be done in the first instance), using 2 cm. of the spectrograph slit, the uncertainty could be reduced to only 0·01 %, and therefore ignored by comparison with other measuring errors. Analysis shows that the uncertainty could then be reduced by 30 %, so that, taken in conjunction with the increased number of comparisons, the standard error becomes ± 0·00007 A, thereby enabling the fourth decimal place to be given with some confidence.

The above modification still requires a fixed reference line upon the spectrograph slit. This might conveniently be produced by partial reflexion from the quartz window attached to the front of the echelon mount, or illuminating a polished horizontal wire in the plane of the spectrograph slit, so that the reference line and echelon fringe have comparable intensity. Thus fractional parts are still obtained by measuring the distance of a fringe from an image of a fixed point upon the spectrograph slit but, compared with the present method, only two-thirds of the number of settings would be required—an important consideration when extensive work is on hand.

The proposed method avoids a difficulty inherent to the echelon optical system, whereby different parts of the echelon lens are used by the echelon and the reference mirrors. Further, when rays are traced after the formation of the image upon the spectrograph slit, it is found that rays from the echelon and reference mirrors use different parts of the spectrograph lens and prism. Thus the three sets of rays pursue rather different optical paths, although it is most desirable that they should receive exactly similar treatment.

This work was carried out in the Physics Department of King's College, Newcastle-upon-Tyne, under the direction of Professor W. E. Curtis, F.R.S., to whom I wish to express my grateful appreciation of his encouragement and help. The reflecting echelon was purchased with the assistance of a grant from the Rockefeller Foundation.

REFERENCES

Barrell, H. & Sears, J. E. Jr. 1939 *Phil. Trans.* A, **238**, 1.
Buisson & Fabry 1913 *C.R. Acad. Sci., Paris*, **156**, 945.
Drinkwater, Richardson & Williams 1940 *Proc. Roy. Soc.* A, **174**, 169.
Humphreys 1931 *Bur. Stand. J. Res.* **7**, 453.
Jackson, C. V. 1936 *Phil. Trans.* A, **236**, 1.
Kopfermann & Wieth Knudsen 1933 *Z. Phys.* **85**, 353.
Kösters & Lampe 1934 *Phys. Z.* **35**, 223.
Meggers, Bruin & Humphreys 1929 *Bur. Stand. J. Res.* **3**, 129.
Pérard 1934 *Trav. Bur. Int. Poids Mes.* **19**, 78.
Williams, W. E. 1933 *Proc. Phys. Soc.* **45**, 699.
Williams, W. E. & Middleton, A. 1939 *Proc. Roy. Soc.* A, **172**, 159.

The distribution of stress in the neighbourhood of a crack in an elastic solid

By I. N. Sneddon, *Bryce Fellow of the University of Glasgow*

(*Communicated by N. F. Mott, F.R.S.—Received* 23 *October* 1945)

The distribution of stress produced in the interior of an elastic solid by the opening of an internal crack under the action of pressure applied to its surface is considered. The analysis is given for 'Griffith' cracks (§ 2) and for circular cracks (§ 3), it being assumed in the latter case that the applied pressure varies over the surface of the crack. For both types of crack the case in which the pressure is constant over the entire crack surface is considered in some detail, the stress components being tabulated and the distribution of stress shown graphically. The effect of a crack (of either type) on the stress produced in an elastic body by a uniform tensile stress is considered and the conditions for rupture deduced.

1. Introduction

1·1. A theory of rupture of non-ductile materials such as glass has been put forward by A. A. Griffith (1920, 1924) on the basis of the existence of a large number of cracks in the interior of the solid body. The fundamental concept of the Griffith theory is that the bounding surfaces of a solid possess a surface tension, just as those of a liquid do, and that when a crack spreads the decrease in the strain energy is balanced by an increase in the potential energy due to this surface tension. The calculation of the effect of the presence of a crack on the energy of an elastic body is based on Inglis's solution (Inglis 1913*a*, *b*) of the two-dimensional equations of elastic equilibrium in the space bounded by two concentric ellipses, the crack being then taken to be an ellipse of zero eccentricity. Denoting the surface tension of the material of the solid body by T, the width of the crack by $2c$, and the Young's modulus of the material of the body by E, Griffith showed that, in the case of plane stress, the crack will spread when the stress P, applied normally to the direction of the crack, exceeds the critical value

$$P_c = \sqrt{\frac{2ET}{\pi c}}. \tag{1·1·1}$$

The use of the equations of elastic equilibrium restricts the analysis to materials which obey Hooke's law fairly closely. This will not be true in general, as the high concentration of stress at the edges of the crack will induce a certain amount of plastic flow. Because of the smallness of the region to which the plastic flow is confined the distribution of stress at points remote from the edges of the crack will not be affected and the strain energy will not differ appreciably from that derived from the solution of the elastic equations. A formula differing from (1·1·1) by a factor 0·8 has been derived by Orowan (1934) from rather similar assumptions. In the case of plane strain the formula (1·1·1) is replaced by*

$$P_c = \sqrt{\frac{2ET}{\pi c(1-\sigma^2)}}. \tag{1·1·2}$$

1·2. Griffith's theory of rupture has recently been extended to three dimensions by Sack (1946). Observing that the length of internal cracks does not greatly exceed their width, Sack calculates the conditions of rupture for a solid containing a plane crack bounded by a circle—a 'penny-shaped' crack—when one of the principal stresses is acting normally to the plane of the crack. By treating the crack as an oblate spheroid whose elliptic section has zero eccentricity Sack establishes that rupture will occur when the tensile stress P normal to the crack exceeds the critical value

$$P_c = \sqrt{\frac{\pi ET}{2c(1-\sigma^2)}}, \tag{1·2·1}$$

where c is the radius of the crack, and, as before, T denotes the surface tension of the material of the body, and E its Young's modulus.

The three-dimensional model introduced by Sack thus gives a critical tensile stress differing from the Griffith value (1·1·1) by a factor $\pi/\{2(1-\sigma^2)^{\frac{1}{2}}\}$, and from (1·1·2) by a factor $\frac{1}{2}\pi$.

1·3. In the present paper main consideration will be given to the distribution of stress in the neighbourhood of a circular crack of the type considered by Sack. As is usual it is assumed that the crack is of such dimensions that its depth exceeds the radius of molecular action at all points not very near its edges. The mathematical theory of elasticity then gives the components of stress accurately at all points of the elastic body other than those in the immediate vicinity of the edges of the crack. If the crack is sufficiently large, the error incurred in the calculation of the strain energy is then negligible, for the stress near the edge of the crack is proportional to $r^{-\frac{1}{2}}$, so that the strain energy in any small sphere whose centre coincides with the edge is finite and negligible in comparison with the strain energy in the rest of the solid. The elastic body is assumed to consist of homogeneous isotropic material and to be so large that its dimensions are very much greater than the radius of the crack; the crack may then be considered to be situated in the interior of an infinite elastic medium. In the first instance the elastic body is considered to be deformed by an internal pressure acting across the surfaces of the crack, the effect of tensile stresses applied to a body with an internal crack free from stress being deduced later.

---

* This is given incorrectly in Griffith (1924) as

$$P_c = \sqrt{\left[\frac{2ET}{\pi c}(1-\sigma^2)\right]}.$$

To afford a comparison with the three-dimensional case the distribution of stress in the neighbourhood of a crack in two dimensions (plane strain) is first considered (§ 2); this corresponds to the two-dimensional Griffith model. The elastic body is supposed to be deformed by the opening of the crack under the action of a uniform hydrostatic pressure acting over its surface. The calculations are based on a solution of the elastic equations given by Westergaard (1939); analytical expressions for the components of stress and for the principal shearing stress are derived from Westergaard's stress function. The advantage of the analysis given here is not only that it is simpler than Inglis's but that Cartesian co-ordinates are employed throughout, thus facilitating the interpretation of the results. To give some idea of the distribution of stress in the body the intensity of the principal shearing stress is computed at a network of points in the two-dimensional co-ordinate plane and the system of 'isochromatic' lines (contours of equal principal shearing stress) plotted in the neighbourhood of the crack. The solution corresponding to the opening of an internal crack under the influence of a tensile stress applied to the surface of the body is then indicated and the Griffith criterion (1·1·2) derived.

In the case of a circular crack (§ 3) it is first of all assumed that the internal pressure is a function of the distance from the centre of the crack. Sack's calculations are based on Neuber's solution (Neuber 1934) of the equations of elastic equilibrium in oblate spheroidal co-ordinates. This method could no doubt be adapted to the case of a variable internal pressure but would probably be rather laborious. Even in the case of constant applied pressure considered by Sack the expressions for the components of stress do not lend themselves readily to computation; and, as in the case of Inglis's analysis, the choice of co-ordinate system makes the interpretation of the results somewhat difficult. In the analysis given below cylindrical polar co-ordinates are used. The solution of the equations of elastic equilibrium in these co-ordinates by the method of Hankel transforms, developed recently by Harding & Sneddon (1945), is employed to reduce the problem to that of the solution of a pair of dual integral equations. A relation giving the shape of the crack in terms of the pressure applied to its surface is derived from the known solution of this pair of equations. The converse problem, that of determining the distribution of internal pressure necessary to preserve a given shape of crack, is also considered.

The distribution of stress in the case in which the applied internal pressure remains constant over the entire surface of the crack is considered in some detail. It is first shown that in this instance the crack, assumed to be originally an infinitely thin crevice, has, after the application of the pressure, the shape of a very flat ellipsoid of revolution. The components of stress are calculated at various points in the interior of the elastic body, and the results given in a set of tables which enable the stress components at any point to be obtained by interpolation; the variation of the various stress components in certain planes parallel to the crack is also illustrated graphically. The principal shearing stress is tabulated in a similar fashion and the contours of equal principal shearing stress plotted to show the distribution of stress in the neighbourhood of the crack.

Finally, the effect of a circular crack, free from stress, on the distribution of stress in an elastic body subjected to a uniform tensile stress is considered, and Sack's criterion (1·2·1) derived on energy considerations.

1·4. Since an exact analysis of the stresses near the base of a crack using a model of the kind introduced by Griffith and Sack would be a necessary preliminary to the consideration of such theoretical problems as the determination of the velocity of propagation of cracks, it is perhaps of some interest to quote here expressions for the components of stress in the vicinity of the edge of the crack. In the case of a two-dimensional crack, if the origin of co-ordinates is chosen to be at the edge of the crack and take axes of co-ordinates (ξ, η) along the axis of the crack and perpendicular to that direction, then the shape of the crack is given approximately by the formula

$$\eta^2 = -b\xi. \tag{1·4·1}$$

Taking polar co-ordinates defined by the relations

$$\xi = \delta \cos \psi, \quad \eta = \delta \sin \psi,$$

then in the immediate vicinity of the crack the components of stress are given by the expressions

$$\sigma_x = p_0 \left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \left(\tfrac{3}{4} \cos \tfrac{1}{2}\psi + \tfrac{1}{4} \cos \tfrac{5}{2}\psi\right), \tag{1·4·1}$$

$$\sigma_y = p_0 \left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \left(\tfrac{5}{4} \cos \tfrac{1}{2}\psi - \tfrac{1}{4} \cos \tfrac{5}{2}\psi\right), \tag{1·4·2}$$

$$\tau_{xy} = p_0 \left(\frac{c}{8\delta}\right)^{\frac{1}{2}} \sin \psi \cos \tfrac{3}{2}\psi. \tag{1·4·3}$$

It follows immediately from these formulae that the principal shearing stress is given by the equation

$$\tau = p_0 \left(\frac{c}{8\delta}\right)^{\frac{1}{2}} \sin \psi. \tag{1·4·4}$$

The most striking feature of the analysis in the three-dimensional case is that the expressions for the components of stress in the neighbourhood of the crack differ from those of the two-dimensional case by a numerical factor only. The stresses are derived from (1·4·1–3) by replacing $\sigma_x, \sigma_y, \tau_{xy}$ on the left-hand sides by $\frac{1}{2}(\pi\sigma_r)$, $\frac{1}{2}(\pi\sigma_2)$, $\frac{1}{2}(\pi\tau_{2r})$, where δ and ψ have a similar interpretation in three dimensions. The hoop stress σ_θ has no analogue in two dimensions; it is given by the relation

$$\sigma_\theta = \frac{4\sigma p_0}{\pi} \left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \cos \tfrac{1}{2}\psi. \tag{1·4·5}$$

2. THE TWO-DIMENSIONAL MODEL

2·1. It has been shown by Westergaard (1939)* that the stress function

$$Z = p_0\left(\frac{z}{\sqrt{(z^2-c^2)}}-1\right), \quad (z = x+iy), \tag{2·1·1}$$

describes the distribution of stress in the interior of an infinite 'two-dimensional' elastic medium produced by the opening of an internal crack from $z=-c$ to $z=c$ (c being real) under the action of a uniform liquid pressure p_0 in the crack as the only load. If the value of the normal component of the displacement, u_y, be denoted by w when $y=0$, then it can readily be shown that for $|x|<c$,

$$\left(\frac{x}{c}\right)^2+\left(\frac{w}{\epsilon}\right)^2 = 1, \tag{2·1·2}$$

where the value of the semi-minor axis ϵ is given by the relation

$$\epsilon = 2(1-\sigma^2)p_0c/E. \tag{2·1·3}$$

Equation (2·1·2) shows that as a result of the elastic strain the shape of the internal crack is elliptic. Furthermore, equation (2·1·1) shows that w is zero for $|x|>c$.

If one adopts the notation

$$z = re^{i\theta}, \quad z-c = r_1e^{i\theta_1}, \quad z+c = r_2e^{i\theta_2}, \tag{2·1·4}$$

then the components of stress are given by the equations

$$\left.\begin{aligned}\tfrac{1}{2}(\sigma_x+\sigma_y) &= \quad ReZ \quad = p_0\left\{\frac{r}{r_1^{\frac{1}{2}}r_2^{\frac{1}{2}}}\cos\left(\theta-\tfrac{1}{2}\theta_1-\tfrac{1}{2}\theta_2\right)-1\right\},\\ \tfrac{1}{2}(\sigma_y-\sigma_x) &= \quad yImZ' = p_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}\sin\tfrac{3}{2}(\theta_1+\theta_2),\\ \tau_{xy} &= -yReZ' = p_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}\cos\tfrac{3}{2}(\theta_1+\theta_2).\end{aligned}\right\} \tag{2·1·5}$$

By means of these expressions the components of stress at any point in the medium can be calculated; the maximum shearing stress across any plane through the point (x, y) can then be obtained from the relation

$$\tau^2 = (\tfrac{1}{2}\sigma_y-\tfrac{1}{2}\sigma_x)^2+\tau_{xy}^2. \tag{2·1·6}$$

Substituting from (2·1·5) into (2·1·6) then, for the maximum shearing stress,

$$\tau = p_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}. \tag{2·1·7}$$

---

* *Note added in proof.* The solution of the elastic equations due to Westergaard, and used in §2 of this paper, corresponds to a uniform distribution of pressure along the surfaces of the crack. By using a method similar to that of §3, but involving Fourier cosine transforms instead of Hankel transforms, it is possible to extend the analysis to cases in which the pressure along the crack is variable. This has been done and will be published shortly in *The Quarterly of Applied Mathematics* by Sneddon, I. N. & Elliott, H. A. under the title: 'The opening of a Griffith crack under internal pressure.'

It is easily verified from the equations (2·1·5) that when $y = 0$ and $-c<x<c$

$$\sigma_x = \sigma_y = -p_0, \quad \tau_{xy} = 0,$$

and that when $|x|>c$

$$\sigma_x = \sigma_y = p_0\left(\frac{x}{\sqrt{(c^2-x^2)}}-1\right), \quad \tau_{xy} = 0.$$

Thus if $x-c$ is small and positive the normal component of stress $[\sigma_y]_{y=0}$ is given approximately by the expression

$$[\sigma_y]_{y=0} \doteqdot p_0\sqrt{\frac{c}{2(x-c)}}. \tag{2·1·8}$$

To determine the distribution of stress in the immediate vicinity of the crack r_1 is taken to be a small quantity, δ say; to conform with the notation of § 1·4 take θ_1 to be ψ, but it should be noted that there are no restrictions on the magnitude of ψ. Since δ/c is assumed small then, from (2·1·4),

$$\left.\begin{aligned} r = c+\delta\cos\psi, \quad r_2 = 2c+\delta\cos\psi,\\ \theta = \delta\sin\psi/c, \quad \theta_2 = \delta\sin\psi/2c, \end{aligned}\right\} \tag{2·1·9}$$

to the first order of small quantities. Substituting from (2·1·9) into equations (2·1·5) then the expressions

$$\tfrac{1}{2}(\sigma_x+\sigma_y) = p_0\left[\left(\frac{c}{2\delta}\right)^{\frac{1}{2}}\cos\tfrac{1}{2}\psi\left\{1-\frac{3\delta}{4c}+O(\delta^2/c^2)\right\}-1\right], \tag{2·1·10}$$

$$\tfrac{1}{2}(\sigma_y-\sigma_x) = p_0\left(\frac{c}{8\delta}\right)^{\frac{1}{2}}\sin\psi\left\{\sin\tfrac{3}{2}\psi-\frac{3\delta}{4c}\sin\tfrac{1}{2}\psi+O(\delta^2/c^2)\right\}, \tag{2·1·11}$$

$$\tau_{xy} = p_0\left(\frac{c}{8\delta}\right)^{\frac{1}{2}}\sin\psi\left\{\cos\tfrac{3}{2}\psi-\frac{3\delta}{4c}\cos\tfrac{1}{2}\psi+O(\delta^2/c^2)\right\}, \tag{2·1·12}$$

are obtained for the determination of the components of stress near the edge of the crack. Similarly, one obtains from equation (2·1·7)—or directly from equations (2·1·11) and (2·1·12)—the expression

$$\tau = p_0\left(\frac{c}{8\delta}\right)^{\frac{1}{2}}\sin\psi\left\{1-\frac{3\delta}{4c}\cos\psi+O(\delta^2/c^2)\right\} \tag{2·1·13}$$

for the principal shearing stress. If terms in δ/c of order greater than $-\frac{1}{2}$ are neglected in these expressions then one obtains the approximate formulae (1·4·1–3). Again, putting $x = c+\xi$, $w = \eta$, $b = 2\epsilon^2/c$ in equation (2·1·2) then the approximate relation

$$\eta^2 = -b\xi \tag{2·1·14}$$

is obtained on neglecting terms of the second order in ξ.

2·2. To give some idea of the distribution of stress in the neighbourhood of the crack, the maximum shearing stress τ was calculated for several values of x and y by means of equation (2·1·7). The results are embodied in table 1 and the variation of τ with x and y shown graphically in figure 1. In the case of plane strain (which is the case considered here) the condition of constant energy of distortion states (Nadai 1931, p. 184) that plastic flow is initiated when τ reaches the value k, where the constant k is related to s_0, the yield point in tension of the material, by the equation

$$3k^2 = s_0^2. \tag{2·2·1}$$

The maximum shear theory gives precisely the same condition except that now

$$2k = s_0. \tag{2·2·2}$$

Thus in both cases the inception of plastic flow can be predicted from the behaviour of the function τ defined by equation (2·1·7).

TABLE 1. VARIATION OF τ/p_0 WITH x AND y IN THE TWO-DIMENSIONAL PROBLEM

x y	0·0	0·2	0·4	0·6	0·8	1·0	1·2	1·4	1·6	1·8	2·0	3·0
0·2	0·189	0·199	0·236	0·327	0·546	0·785	0·405	0·179	0·094	0·058	0·037	0·009
0·4	0·320	0·332	0·372	0·444	0·534	0·543	0·400	0·248	0·153	0·099	0·068	0·017
0·6	0·378	0·386	0·408	0·439	0·456	0·428	0·346	0·252	0·176	0·124	0·089	0·024
0·8	0·381	0·384	0·390	0·395	0·386	0·354	0·298	0·235	0·178	0·134	0·101	0·031
1·0	0·354	0·354	0·352	0·345	0·329	0·299	0·262	0·210	0·171	0·135	0·106	0·036
1·2	0·315	0·313	0·308	0·298	0·281	0·256	0·225	0·192	0·159	0·130	0·106	0·039
1·4	0·275	0·273	0·267	0·257	0·242	0·222	0·198	0·172	0·147	0·124	0·103	0·042
1·6	0·238	0·236	0·231	0·222	0·209	0·193	0·174	0·154	0·134	0·115	0·099	0·044
1·8	0·206	0·205	0·200	0·192	0·182	0·169	0·154	0·138	0·123	0·107	0·093	0·044
2·0	0·179	0·178	0·174	0·168	0·159	0·149	0·137	0·124	0·112	0·099	0·087	0·044
3·0	0·095	0·094	0·093	0·091	0·088	0·084	0·080	0·076	0·071	0·066	0·061	0·039

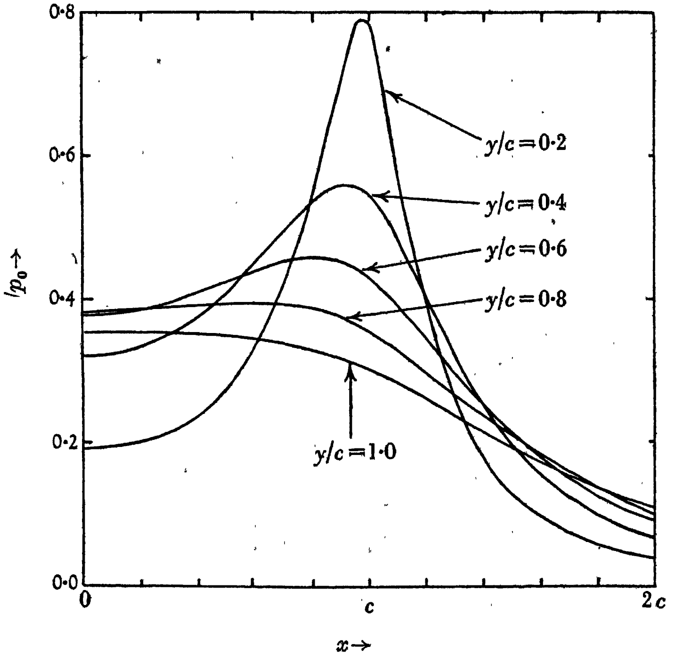


FIGURE 1. The variation of the principal shearing stress, τ, with x and y.

A convenient method of showing the variation of this function and of visualizing the distribution of stress in the interior of the medium consists of plotting the contours of equal principal shearing stress, i.e. constructing the family of curves $\tau = \alpha p_0$, or

$$c^2y^4 = \alpha\{(x-c)^2+y^2\}^{\frac{3}{2}}\{(x+c)^2+y^2\}^{\frac{3}{2}}, \tag{2·2·3}$$

where α is a parameter. These curves are the 'isochromatic lines' of photoelasticity. The family of curves was constructed by determining from table 1 and figure 1 the points in the vicinity of the crack at which the parameter α has the values 0·7, 0·5, 0·4, 0·3 and 0·2. The result is shown in figure 2. The shape of the isochromatic lines in the neighbourhood of the edge of the crack can readily be deduced from equation (2·1·13); in polar co-ordinates δ and ψ with origin at $x = c$ the equation of an isochromatic curve reduces to the simple form

$$\delta = \beta \sin^2 \psi, \tag{2·2·4}$$

where $\beta = c/8\alpha^2$. Thus near the edge of the crack the isochromatics are curves of two loops situated symmetrically with regard to the axis of the crack.

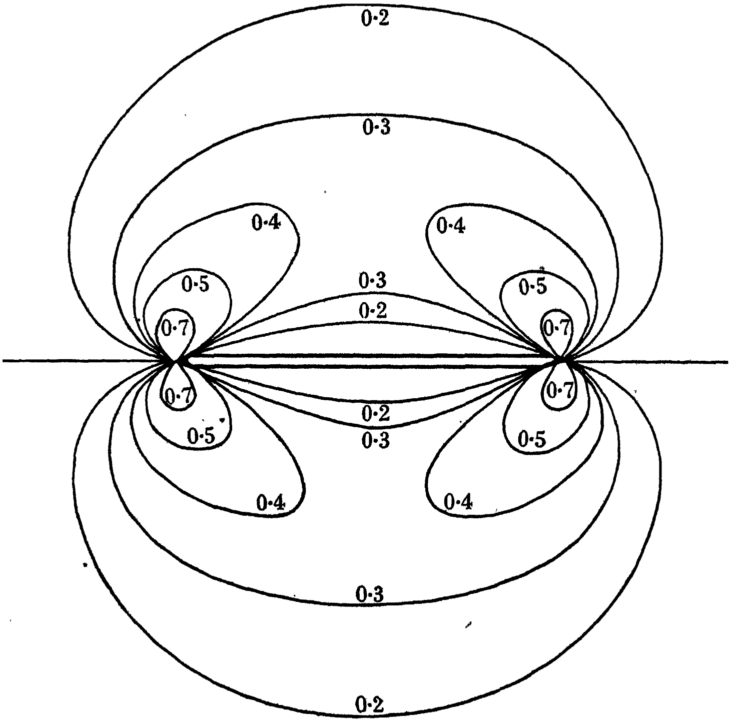


Figure 2. The isochromatic lines in the vicinity of a Griffith crack.

The fact that all the isochromatic lines pass through the points $(\pm c, 0)$ shows that at both of these points the principal shearing stress is infinite; thus even for small internal pressures p_0 plastic flow occurs at the corners of the crack to remove this infinite stress. There is, in fact, no purely *elastic* solution of the problem; if, however, the internal pressure is not too large the region of plastic flow will be small and will not appreciably affect the distribution of stress at points in the solid at a distance from the corners of the crack.

Although there seems to be no photoelectric determinations of the stress distribution in the vicinity of a crack there is a certain degree of similarity between the theoretical isochromatic lines of figure 2 and those in the vicinity of a tunnel as determined experimentally by Farquharson (Frocht 1941, p. 347). The shape of the tunnel differs appreciably from that of a crack of small transverse thickness, but the resemblance of the shape of the isochromatic lines in the neighbourhood of the lower straight edge of the tunnel to that of the lines of figure 2 is striking.

2·3. Now consider the energy of the crack. When the semi-minor axis of the crack is ϵ the internal pressure has the value

$$p = \frac{E}{2(1-\sigma^2)}\left(\frac{\epsilon}{c}\right).$$

Considering an element of length of the crack it is found that the work done in increasing the semi-minor axis to $\epsilon + d\epsilon$ is

$$2p\,dx\,dw = \frac{E}{1-\sigma^2}\left(\frac{\epsilon+\frac{1}{2}d\epsilon}{c}\right)dx\sqrt{\left(1-\frac{x^2}{c^2}\right)}\,d\epsilon,$$

so that the energy of the crack is

$$\begin{aligned} W &= \frac{2E}{(1-\sigma^2)\,c}\int_0^c\sqrt{\left(1-\frac{x^2}{c^2}\right)}\,dx\int_0^\epsilon \epsilon\,d\epsilon \\ &= \frac{\pi E\epsilon^2}{4(1-\sigma^2)} = \frac{\pi(1-\sigma^2)\,p_0^2c^2}{E}, \qquad (2\cdot3\cdot1) \end{aligned}$$

where p_0 refers to the final pressure.

The theory of the Griffith crack, considers the distribution of stress in the vicinity of a crack when the crack is opened under the influence of an average tension p_0 applied to the surface of the body; the surface of the crack is assumed to be free from stress. The solution of this problem is obtained by superposing on the solution (2·1·1) a second solution

$$Z = +p_0. \qquad (2\cdot3\cdot2)$$

This ensures that $\sigma_x = \sigma_y = p_0$ for large values of x and y, and that $\sigma_x = \tau_{xy} = 0$ across the surface of the crack. Equation (2·3·1) then shows that the presence of a crack of length $2c$ lowers the potential energy of the solid by an amount W. On the other hand, the crack has a surface energy

$$U = 4cT, \qquad (2\cdot3\cdot3)$$

where T is the surface tension of the material. Thus the total diminution of the potential energy due to the presence of the crack is

$$W - U = \pi(1-\sigma^2)\,p_0^2c^2/E - 4cT.$$

The condition $$\frac{\partial}{\partial c}(W-U)=0$$

that the crack may extend then leads to the equation

$$c=\frac{2ET}{\pi p_0^2(1-\sigma^2)}.$$

If c is less than this value or if p_0 exceeds the critical value p_c given by the relation

$$p_c=\sqrt{\left\{\frac{2ET}{\pi c(1-\sigma^2)}\right\}}, \tag{2·3·4}$$

the crack will become unstable and spread. Equation (2·3·4) is the (corrected) Griffith criterion for rupture in the case of plane strain, equation (1·1·2) above.

3. THE THREE-DIMENSIONAL MODEL

3·1. In the three-dimensional case assume that a crack is created in the interior of an infinite elastic medium and that it is 'penny-shaped', occupying the circle $r^2=x^2+y^2=c^2$ in the plane $z=0$. An examination of the equations of § 2·1 shows immediately that in the case of the two-dimensional model the distribution of stress in the neighbourhood of a crack of length $2c$ lying along the axis $y=0$, is identical to that in the semi-infinite elastic medium $y \geqslant 0$ when the conditions on the boundary $y=0$ are prescribed by the relations

$$\left.\begin{aligned}&\text{(i)}\quad \tau_{xy}=0 \quad \text{for all values of } x,\\ &\text{(ii)}\quad \sigma_y=-p_0\ \ (|x|<c),\quad u_y=0\ \ (|x|>c).\end{aligned}\right\} \tag{3·1·1}$$

In the three-dimensional case assume that the distribution of stress in the neighbourhood of the crack is the same as that produced in a semi-infinite elastic medium when certain conditions are prescribed on its bounding plane.

For a crack of the shape described above there is symmetry about the z-axis, so cylindrical co-ordinates r, θ, z may be employed; in this co-ordinate system the displacement vector assumes the form $(u_r, 0, u_z)$, and the stress in the interior of the medium is specified completely by the stress components $\sigma_r, \sigma_z, \sigma_\theta, \tau_{rz}$, the remaining components being identically zero.

By analogy with equations (3·1·1) it may now be assumed that the distribution of stress in the neighbourhood of the crack is the same as that produced in the interior of the semi-infinite elastic solid $z \geqslant 0$ by the boundary conditions

$$\left.\begin{aligned}&\text{(i)}\quad \tau_{rz}=0 \quad \text{for all values of } r,\\ &\text{(ii)}\quad \sigma_z=-p(r)\ \ (r<c),\quad u_z=0\ \ (r>c),\end{aligned}\right\} \tag{3·1·2}$$

on the plane $z=0$. In the first instance assume that the applied internal pressure p is a function of r; the case of a constant internal pressure will be examined later.

The only restriction on the function $p(r)$ imposed by the subsequent analysis is that it is such that the integral

$$\int_0^c rp(r)\,J_0(\xi r)\,dr$$

exists for all real values of the parameter ξ.

The conditions (3·1·2) do not correspond to a hydrostatic pressure in the crack in all cases. In the analyses of Griffiths and Sack the crack is assumed to be of very small depth, so that the surface of the crack may be taken to be coincident with the z-axis; in this instance the pressure across the surface of the crack is $-\sigma_z$ and the conditions (3·1·2) are exact. Even in cases where the depth of the crack is taken into account the distribution of stress given by the conditions (3·1·2) will not differ much from that given by a hydrostatic pressure inside the crack—except possibly near the boundary of the crack. It will be seen later (§ 3·5) that in the case where the internal pressure is constant throughout the crack and the medium is incompressible the conditions (3·1·2) correspond exactly to a hydrostatic pressure.

With the axial symmetry the equations of elastic equilibrium reduce in the absence of body forces to (Love 1934, p. 90)

$$\frac{\partial\sigma_r}{\partial r}+\frac{\partial\tau_{rz}}{\partial z}+\frac{\sigma_r-\sigma_\theta}{r}=0, \tag{3·1·3}$$

$$\frac{\partial\tau_{rz}}{\partial r}+\frac{\partial\sigma_z}{\partial z}+\frac{\tau_{rz}}{r}=0. \tag{3·1·4}$$

It has been shown (Harding & Sneddon 1945) that these equations and the conditions of compatibility are satisfied by the forms

$$\sigma_r=\int_0^\infty \xi\{\lambda G'''+(\lambda+2\mu)\,\xi^2G'\}\,J_0(\xi r)\,d\xi-\frac{2(\lambda+\mu)}{r}\int_0^\infty \xi^2G'J_1(\xi r)\,d\xi, \tag{3·1·5}$$

$$\sigma_z=\int_0^\infty \xi\{(\lambda+2\mu)\,G'''-(3\lambda+4\mu)\,\xi^2G'\}\,J_0(\xi r)\,d\xi, \tag{3·1·6}$$

$$\sigma_\theta=\lambda\int_0^\infty \xi\{G'''-\xi^2G'\}\,J_0(\xi r)\,d\xi+\frac{2(\lambda+\mu)}{r}\int_0^\infty \xi^2G'J_1(\xi r)\,d\xi, \tag{3·1·7}$$

$$\tau_{rz}=\int_0^\infty \xi^2\{\lambda G''+(\lambda+2\mu)\,\xi^2G\}\,J_1(\xi r)\,d\xi, \tag{3·1·8}$$

and as noted above $\tau_{z\theta}=\tau_{r\theta}=0.$ (3·1·9)

In these expressions λ and μ are Lamé's elastic constants and the function $G(\xi,z)$ is a solution of the ordinary differential equation

$$\left(\frac{d^2}{dz^2}-\xi^2\right)^2G=0. \tag{3·1·10}$$

The constants of integration introduced into the solution of this fourth-order equation are determined from the imposed boundary conditions. The solutions given above are such that all the components of stress tend to zero as $r\to\infty$; to ensure further that all these components tend to zero as $z\to\infty$ assume a solution of equation (3·1·10) of the form

$$G = (A+Bz)\,e^{-\xi z}, \tag{3·1·11}$$

where A and B are functions of the parameter ξ.

The corresponding expressions for the non-vanishing components of the displacement vector are

$$u_r = \frac{\lambda+\mu}{\mu}\int_0^\infty \xi^2 G' J_1(\xi r)\,d\xi, \tag{3·1·12}$$

$$u_z = \int_0^\infty \xi\left\{G'' - \frac{\lambda+2\mu}{\mu}\xi^2 G\right\} J_0(\xi r)\,d\xi. \tag{3·1·13}$$

To determine the values of the arbitrary functions A and B of equation (3·1·11) it is necessary to evaluate τ_{rz}, u_z, σ_z on the plane $z=0$. By repeated differentiation of (3·1·11) and by use of the results

$$\lambda = \frac{E\sigma}{(1+\sigma)(1-2\sigma)}, \quad \mu = \frac{E}{2(1+\sigma)} \tag{3·1·14}$$

expressing the Lamé constants in terms of the Young's modulus E and the Poisson ratio σ, then, on substituting $z=0$ into equation (3·1·8),

$$[\tau_{rz}]_{z=0} = \frac{E}{(1+\sigma)(1-2\sigma)}\int_0^\infty \xi^3(\xi A - 2\sigma B)\,J_1(\xi r)\,d\xi.$$

Inverting this result by means of the Hankel inversion theorem, then

$$\xi^2(\xi A - 2\sigma B) = \frac{(1+\sigma)(1-2\sigma)}{E}\int_0^\infty r[\tau_{rz}]_{z=0}\,J_1(\xi r)\,d\xi.$$

If the boundary conditions $[\tau_{rz}]_{z=0}=0$ holds for all values of r it follows that

$$\xi A = 2\sigma B \tag{3·1·15}$$

between the two constants of integration A and B. Substituting from equation (3·1·15) into equation (3·1·11), differentiating with respect to z and then putting $z=0$ it is found from equations (3·1·6) and (3·1·13) that

$$[\sigma_z]_{z=0} = \frac{E}{(1+\sigma)(1-2\sigma)}\int_0^\infty \xi^3 B J_0(\xi r)\,d\xi, \tag{3·1·16}$$

$$[u_z]_{z=0} = -\frac{2(1-\sigma)}{1-2\sigma}\int_0^\infty \xi^2 B J_0(\xi r)\,d\xi. \tag{3·1·17}$$

Writing (3·1·11) in the form

$$G = \frac{B(c\xi)}{\xi}(2\sigma+\xi z)\,e^{-\xi z} \tag{3·1·18}$$

and then putting $\eta = \xi c$ in equations (3·1·16) and (3·1·17), one obtains finally

$$[\sigma_z]_{z=0} = \frac{E}{c^4(1+\sigma)(1-2\sigma)} \int_0^\infty \eta f(\eta)\, J_0(\rho\eta)\, d\eta \qquad (3\cdot1\cdot19)$$

and
$$[u_z]_{z=0} = -\frac{2(1-\sigma)}{c^3(1-2\sigma)} \int_0^\infty f(\eta)\, J_0(\rho\eta)\, d\eta, \qquad (3\cdot1\cdot20)$$

where
$$f(\eta) = \eta^2 B(\eta) \qquad (3\cdot1\cdot21)$$

and $\rho = r/c$.

3·2. Inserting the boundary conditions

$$z = 0,\; u_z = 0 \;\; (r > c), \quad \sigma_z = -p(r) \;\; (r < c)$$

into equations (3·1·19) and (3·1·20), the pair of integral equations

$$\left.\begin{aligned} &\int_0^\infty \eta f(\eta)\, J_0(\rho\eta)\, d\eta = g(\rho) \quad (0 < \rho < 1),\\ &\int_0^\infty f(\eta)\, J_0(\rho\eta)\, d\eta = 0 \qquad\;\; (\rho > 1), \end{aligned}\right\} \qquad (3\cdot2\cdot1)$$

is obtained for the determination of the unknown function $f(\eta)$ and hence of the constant $B(c\xi)$ of equation (3·1·18); in the former of these two equations is written

$$g(\rho) = -\frac{(1+\sigma)(1-2\sigma)}{E} c^4 p(\rho c). \qquad (3\cdot2\cdot2)$$

Dual integral equations of the type (3·2·1) have been considered by Titchmarsh (1937, p. 334) and Busbridge (1938); the solution of the pair (3·2·1) can be derived easily from Busbridge's analysis in the form

$$f(\eta) = \frac{2}{\pi} \int_0^1 \mu \sin \mu\eta\, d\mu \int_0^1 \frac{\rho g(\rho\mu)}{\sqrt{(1-\rho^2)}}\, d\rho. \qquad (3\cdot2\cdot3)$$

Substituting from (3·2·3) into (3·1·20) and making use of the result (Watson 1944, p. 405)

$$\int_0^\infty \sin \mu\eta J_0(\rho\eta)\, d\eta = \begin{cases} 0 & (\rho > \mu), \\ (\mu^2 - \rho^2)^{-\frac{1}{2}} & (\rho < \mu), \end{cases}$$

it is found that when the applied pressure is $p(r)$ the value of the normal component of the surface displacement is given by the equation

$$[u_z]_{z=0} = \frac{4(1-\sigma^2)\,c}{\pi E} \int_\rho^1 \frac{\mu\, d\mu}{\sqrt{(\mu^2-\rho^2)}} \int_0^1 \frac{x p(x\mu c)}{\sqrt{(1-x^2)}}\, dx \quad (\rho < 1). \qquad (3\cdot2\cdot4)$$

If, for example, the applied pressure is given by the power series

$$p(r) = p_0 \sum_{n=0}^\infty \alpha_n \left(\frac{r}{c}\right)^n, \quad (0 < r < c), \qquad (3\cdot2\cdot5)$$

then substituting into equation (3·2·3) one obtains

$$f(\eta) = -\frac{(1+\sigma)(1-2\sigma)}{\sqrt{\pi} E} c^4 p_0 \sum_{n=0}^\infty \frac{\Gamma(1+\frac{1}{2}n)}{\Gamma(\frac{3}{2}+\frac{1}{2}n)} \alpha_n \int_0^1 \mu^{n+1} \sin \mu\eta\, d\mu, \qquad (3\cdot2\cdot6)$$

so that the normal component of the surface displacement now has the value

$$[u_z]_{z=0} = \frac{2}{\sqrt{\pi}}(1-\sigma^2)\frac{p_0}{E}\sum_{n=0}^{\infty}\frac{\Gamma(1+\frac{1}{2}n)}{\Gamma(\frac{3}{2}+\frac{1}{2}n)}\alpha_n I_n\left(\frac{r}{c}\right)^{n+1} \quad (0<r<c), \tag{3·2·7}$$

where I_n denotes the integral

$$I_n = c\int_1^{c/r}\frac{\eta^{n+1}}{\sqrt{(\eta^2-1)}}\,d\eta. \tag{3·2·8}$$

Integrating by parts it can readily be shown that I_n satisfies the recurrence relation

$$I_n = \frac{n}{n+1}I_{n-2}+\frac{1}{n+1}\left(\frac{c}{r}\right)^{n+1}(c^2-r^2)^{\frac{1}{2}}. \tag{3·2·9}$$

By means of these equations the value of the normal component of the surface displacement can be calculated when the law of variation of the applied pressure is known.

The value of the radial component of the surface displacement is found from equation (3·1·12) to be

$$[u_r]_{z=0} = \frac{1}{c^3}\int_0^{\infty} f(\eta)\,J_1(\rho\eta)\,d\eta, \tag{3·2·10}$$

whence by equations (3·2·2) and (3·2·3) and making use of the result (Watson 1944, p. 405) we obtain

$$\int_0^{\infty} J_1(\rho\eta)\sin\mu\eta\,d\eta = \begin{cases}\frac{\mu}{\rho}(\rho^2-\mu^2)^{-\frac{1}{2}} & (\mu<\rho),\\ 0 & (\mu>\rho),\end{cases} \tag{3·2·11}$$

then, in the general case,

$$[u_r]_{z=0} = -\frac{2(1+\sigma)(1-2\sigma)c}{\pi E}\int_0^1\mu^2 d\mu\int_\mu^1\frac{p(\rho\mu c)\,d\rho}{\sqrt{\{(1-\rho^2)(\rho^2-\mu^2)\}}}. \tag{3·2·12}$$

It will be observed from this equation that the radial component of the surface displacement is zero when the medium is incompressible ($\sigma = \frac{1}{2}$).

3·3. If it be supposed that the applied pressure $p(r)$ is constant over a circular area of radius $a \leqslant c$, then

$$p(r) = \begin{cases}p_0 & (0<r<a),\\ 0 & (a<r<c).\end{cases} \tag{3·3·1}$$

Substituting into equation (3·2·4) one obtains the expression

$$w = \frac{4p_0(1-\sigma^2)}{\pi E}(c^2-r^2)^{\frac{1}{2}}\{1-(1-a^2/c^2)^{\frac{1}{2}}\} \tag{3·3·2}$$

for the normal component of the surface displacement, $w = [u_z]_{z=0}$. If $w = \epsilon$ when $r = 0$ then

$$\epsilon = \frac{4(1-\sigma^2)p_0 c}{\pi E}\{1-(1-a^2/c^2)^{\frac{1}{2}}\}, \tag{3·3·3}$$

and equation (3·3·2) may be written in the form

$$\frac{r^2}{c^2}+\frac{w^2}{\epsilon^2}=1, \tag{3·3·4}$$

showing that for all values of $a \leqslant c$, the crack resulting from the application of an internal pressure is ellipsoidal in shape provided the applied pressure is constant. For a point force applied at the centre of symmetry allow the radius a to tend to zero and the applied pressure p_0 to tend to infinity in such a way that the total force, $\pi a^2 p_0$, remains constant, P say; in this instance one obtains from equation (3·3·3)

$$\epsilon = \frac{2(1-\sigma^2)}{\pi^2}\frac{P}{cE}.$$

If the pressure p_0 acts over a circle of radius c, i.e. over the entire surface of the crack, and produces a displacement ϵ of the centre of the circle then

$$p_0 = \frac{\pi E \epsilon}{4(1-\sigma^2)c}, \tag{3·3·5}$$

from which it will be observed that p_0 is constant for constant values of the ratio ϵ/c. Now consider the work done in forming an ellipsoidal depression of circular section of radius c and of depth ϵ. When the depth of the depression is ϵ the pressure is given by equation (3·3·5); when the depth of the depression has increased to $\epsilon + d\epsilon$ the pressure will have increased to $\pi E(\epsilon + d\epsilon)/[4(1-\sigma^2)c]$, so that in bringing about a change $d\epsilon$ of the depth of the ellipsoidal depression the average pressure is

$$p_0 = \frac{\pi E(\epsilon + \frac{1}{2}d\epsilon)}{4(1-\sigma^2)c}.$$

Considering a ring element of surface intersecting the plane $\theta = 0$ at the points P, P' (see figure 3) it is found that the work done in making a small normal displacement dw is

$$p_0 2\pi r\,dr\,dw = \frac{\pi E(\epsilon + \frac{1}{2}d\epsilon)}{4(1-\sigma^2)c} 2\pi r\,dr\,d\epsilon(1-r^2/c^2)^{\frac{1}{2}}$$

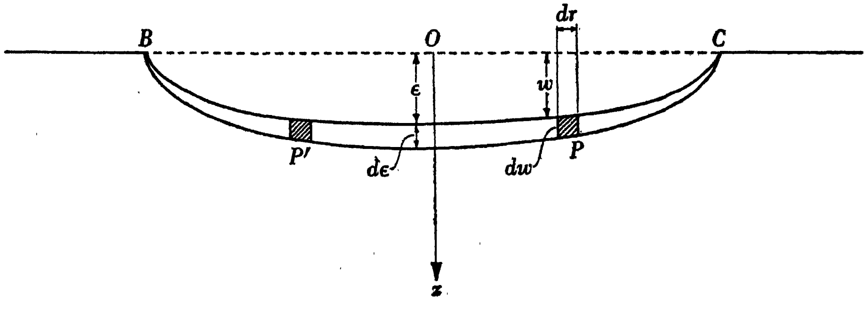


FIGURE 3

by equation (3·3·2). Thus the total work done in forming the ellipsoidal depression on the surface of a semi-infinite elastic solid is

$$W_1 = \frac{\pi^2 E}{2(1-\sigma^2)} \int_0^c r(1-r^2/c^2)^{\frac{1}{2}}\, dr \int_0^{\epsilon} \epsilon\, d\epsilon,$$

which gives

$$W_1 = \frac{\pi^2 E c \epsilon^2}{12(1-\sigma^2)}.$$

The energy of a crack formed by internal pressure is twice this energy; multiplying the last result by 2 and substituting for the value of ϵ in terms of the final pressure p_0 one obtains for the energy of a crack of radius c the formula

$$W = 2W_1 = \frac{8(1-\sigma^2)}{3E} p_0^2 c^3. \qquad (3·3·6)$$

3·4. It was observed above (equations (3·3·3) and (3·3·4)) that as long as the internal applied pressure is constant the crack will be ellipsoidal in shape whatever the radius of the circle over which it is applied. It is natural then to try to determine whether cracks of a shape other than ellipsoidal are possible, and if they are to determine the distribution of internal pressure necessary to preserve their shape. This is equivalent to assuming that the value of $[u_z]_{z=0}$ is known when $r \leqslant c$, but that the function $p(r)$ is unknown. One might then regard equation (3·2·4) as an integral equation determining $p(r)$ when $[u_z]_{z=0}$ is known. It is, however, simpler to consider the problem of determining the distribution of stress in a semi-infinite elastic medium bounded by the plane $z = 0$ when the surface value of the normal component of the displacement vector is prescribed for all values of r. The analysis proceeds along similar lines except that now the boundary conditions

$$\text{(i)} \quad \tau_{zr} = 0, \quad \text{for all values of } r,$$

$$\text{(ii)} \quad u_z = \begin{cases} w(r) & (r < c), \\ 0 & (r > c), \end{cases}$$

replace the set (3·1·2). Substituting the condition (ii) into equation (3·1·17) then

$$\int_0^\infty \xi^2 B J_0(\xi r)\, d\xi = -\begin{cases} -\dfrac{(1-2\sigma)}{2(1-\sigma)} w(r) & (r < c), \\ 0 & (r > c). \end{cases}$$

Inverting this result by means of the Hankel inversion theorem

$$\xi B = -\frac{(1-2\sigma)}{2(1-\sigma)} \int_0^c r w(r)\, J_0(\xi r)\, dr. \qquad (3·4·1)$$

Denoting the integral on the right-hand side by $\overline{w}(\xi)$ then, from equation (3·1·16), it is found that

$$[\sigma_z]_{z=0} = -\frac{E}{2(1-\sigma^2)} \int_0^\infty \xi^2 \overline{w}(\xi)\, J_0(\xi r)\, d\xi. \qquad (3·4·2)$$

Equation (3·4·2) gives the value of the applied pressure on the circle $r \leqslant c$ which must be applied to maintain the prescribed surface displacement.

For example, if it is assumed that

$$w(r) = \epsilon(1-r^2/c^2), \tag{3·4·3}$$

then from equation (3·4·1) it follows that

$$\bar{w}(\xi) = 2\epsilon J_2(c\xi)/\xi^2,$$

and so by equation (3·4·2) that the normal component of the surface stress is given by the expression

$$[\sigma_z]_{z=0} = -\frac{E\epsilon}{1-\sigma^2}\int_0^\infty J_2(c\xi)\, J_0(r\xi)\, d\xi.$$

Using the recurrence relation

$$J_2(c\xi) = J_0(c\xi) - \frac{2}{c}\frac{\partial}{\partial \xi} J_1(c\xi),$$

after an integration by parts it is found that

$$\int_0^\infty J_0(r\xi)\, J_2(c\xi)\, d\xi = \int_0^\infty J_0(r\xi)\, J_0(c\xi)\, d\xi - \frac{2r}{c}\int_0^\infty J_1(r\xi)\, J_1(c\xi)\, d\xi.$$

Now by Gübler's formula (Watson 1944, p. 410) the value of the first integral is found to be

$$\left(\frac{1}{c}\right) {}_2F_1\left[\tfrac{1}{2}, \tfrac{1}{2};\ 1;\ \frac{r^2}{c^2}\right] = \frac{2}{\pi c} K(r/c) \quad (r<c),$$

in the usual notation for elliptic integrals (Jahnke-Emde 1938, p. 85). The second integral could also be evaluated by the Gübler formula, but the hypergeometric series involved in this instance is not readily summed. The integration can be carried out more easily by making use of Neumann's result

$$J_1(\xi r)\, J_1(\xi c) = \frac{1}{\pi}\int_0^\pi J_0(\xi R) \cos\theta\, d\theta, \quad R^2 = c^2 + r^2 - 2cr\cos\theta.$$

Multiplying both sides of this equation by $\exp\{-\xi z\}$ and integrating with respect to ξ from 0 to ∞, then

$$\int_0^\infty e^{-\xi z} J_1(c\xi)\, J_1(r\xi)\, d\xi = \frac{1}{\pi}\int_0^\pi \cos\theta\, d\theta \int_0^\infty J_0(\xi R)\, e^{-\xi z}\, d\xi = \frac{1}{\pi}\int_0^\pi \frac{\cos\theta\, d\theta}{\sqrt{(R^2+z^2)}}.$$

Letting z tend to zero on both sides of the last equation one obtains the required result

$$\int_0^\infty J_1(\xi r)\, J_1(\xi c)\, d\xi = \frac{1}{\pi}\int_0^\pi \frac{\cos\theta\, d\theta}{R}$$

The integral on the right-hand side can easily be evaluated in terms of elliptic integrals; with the usual notation for complete elliptic integrals, then

$$\frac{2r}{c}\int_0^\infty J_1(\xi r)\, J_1(\xi c)\, d\xi = \frac{2\{1+(r/c)^2\}}{\pi(c+r)} K(k) - \frac{2(c+r)}{\pi c^2} E(k),$$

where $$k^2 = 4rc/(r+c)^2.$$

Thus the displacement (3·4·3) is maintained by an applied internal pressure

$$p(r) = -[\sigma_z]_{z=0} = \frac{2E\epsilon}{\pi c(1-\sigma^2)} f(\rho), \tag{3·4·4}$$

where the function $f(\rho)$ is defined by the equation

$$f(\rho) = \frac{1+\rho^2}{1+\rho} K(k) - K(\rho) - (1+\rho) E(k) \tag{3·4·5}$$

and $\rho = r/c$, $k^2 = 4\rho/(1+\rho)^2$. The variation of the function $f(\rho)$ with ρ is shown graphically in figure 4. From this curve it follows that the applied internal force is *compressive* in the body of the crack, but that near the circular edge of the crack the applied force becomes *tensile* until round the edge it is both infinite and tensile. This is precisely the kind of variation which would be predicted on physical grounds for a crack of this shape.

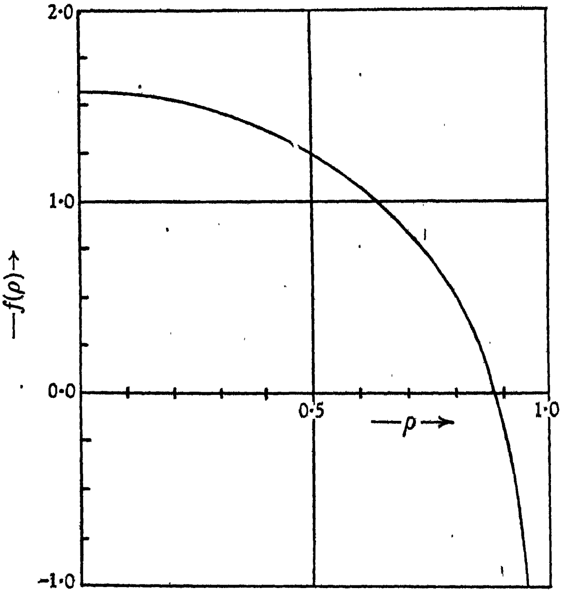


FIGURE 4. Variation of the function $f(\rho)$ with ρ.

3·5. Now return to the problem of determining the distribution of stress in the interior of the medium, in the case where the applied pressure $p(r)$ is a constant p_0 over the entire surface area $r \leqslant c$ of the crack. Putting $p(\rho c) = p_0$, then from equations (3·2·2) and (3·2·3),

$$f(\eta) = \frac{2p_0 c^4 (1+\sigma)(1-2\sigma)}{\pi E} \frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right). \tag{3·5·1}$$

Substituting this value for $f(\eta)$ into equation (3·1·19) it follows that, for the normal component of stress across the plane $z = 0$,

$$[\sigma_z]_{z=0} = \frac{2p_0}{\pi}\left[\rho \int_0^\infty \sin\eta\, J_1(\rho\eta)\, d\eta - \int_0^\infty \frac{\sin\eta}{\eta} J_0(\rho\eta)\, d\eta\right]. \tag{3·5·2}$$

By means of the results (Watson 1944, p. 405)

$$\int_0^\infty \frac{\sin\eta}{\eta} J_0(\rho\eta)\,d\eta = \begin{cases} \sin^{-1}(1/\rho) & (\rho \geqslant 1), \\ \tfrac{1}{2}\pi & (\rho \leqslant 1), \end{cases}$$

$$\int_0^\infty \rho \sin\eta J_1(\rho\eta)\,d\eta = \begin{cases} (\rho^2-1)^{-\frac{1}{2}} & (\rho > 1), \\ 0 & (\rho < 1), \end{cases}$$

it is verified that $\sigma_z = -p_0$ when $\rho < 1$, and for the value of the normal component of the surface stress when $\rho > 1$ the expression

$$[\sigma_z]_{z=0} = -\frac{2p_0}{\pi}\left[\sin^{-1}\frac{1}{\rho} - \frac{1}{\sqrt{(\rho^2-1)}}\right] \tag{3·5·3}$$

is obtained. It will be observed that when $\rho = 1$ the stress $[\sigma_z]_{z=0}$ is infinite and that when $\rho - 1$ is small and positive

$$[\sigma_z]_{z=0} \doteqdot \frac{p_0}{\pi}(\rho-1)^{-\frac{1}{2}}, \tag{3·5·4}$$

showing that there is a large *tensile* stress in the neighbourhood of the circle $\rho = 1$, $z = 0$. Furthermore, as $\rho \to \infty$,

$$[\sigma_z]_{z=0} \sim \frac{4p_0}{3\pi}\rho^{-3}, \tag{3·5·5}$$

the stress being still tensile. Since

$$\frac{d}{d\rho}[\sigma_z]_{z=0} = -\frac{2p_0}{\pi}\{\rho(\rho^2-1)^{\frac{3}{2}}\}^{-1}$$

does not change sign in the region $\rho > 1$, it follows that the normal component of the surface stress $[\sigma_z]_{z=0}$ never becomes compressive in that region. The similarity of the expression (3·5·4) to that for the corresponding stress in the two-dimensional analysis (equation (2·1·8)) is also of interest.

For the values of the other stress components on the plane $z = 0$ one obtains from equations (3·1·5)–(3·1·8) the expressions

$$\begin{aligned} [\sigma_r + \sigma_\theta]_{z=0} &= (1+2\sigma)[\sigma_z]_{z=0}, \\ [\sigma_r - \sigma_\theta]_{z=0} &= \frac{2(1-\sigma)p_0}{\pi}\int_0^\infty \eta J_2(\rho\eta)\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)d\eta \\ &= \frac{2(1-2\sigma)p_0}{\pi}\left[\int_0^\infty J_2(\rho\eta)\cos\eta\,d\eta - \int_0^\infty J_2(\rho\eta)\frac{\sin\eta}{\eta}\,d\eta\right]. \end{aligned}$$

Evaluating the integrals in the square bracket it is found that the bracket vanishes when $\rho < 1$ and has the value

$$-(\rho^2-1)^{-\frac{1}{2}}$$

when $\rho > 1$. Thus if $\rho < 1$ and $z = 0$, then

$$\sigma_\theta = \sigma_r = -(\sigma + \tfrac{1}{2})p_0,$$

and, if $\rho > 1$, $z = 0$, $$\sigma_r = \frac{2p_0}{\pi}\left[(\rho^2-1)^{-\frac{1}{2}} - (\sigma+\tfrac{1}{2})\sin^{-1}\frac{1}{\rho}\right],$$
$$\sigma_\theta = \frac{2p_0}{\pi}\left[2\sigma(\rho^2-1)^{-\frac{1}{2}} - (\sigma+\tfrac{1}{2})\sin^{-1}\frac{1}{\rho}\right].$$

It will be noted that in the case $\sigma = \frac{1}{2}$ (incompressible medium) $\sigma_r = \sigma_z = \sigma_\theta = -p_0$ $(\rho < 1)$, so that in this instance the internal pressure is truly hydrostatic.

3·6. The evaluation of the stress components in the interior of the medium is similar to that involved in the solution of the Boussinesq problem for a cylinder (Sneddon 1946). Obtaining the value of $B(c\xi)$ from equations (2·1·21) and (3·5·1) and substituting from equation (3·1·18) into equations (3·1·12) and (3·1·13), then, for the non-vanishing components of the displacement vector,

$$u_r = \frac{2p_0 c}{\pi E}(1+\sigma)\int_0^\infty (1-2\sigma-\zeta\eta)\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)e^{-\zeta\eta}J_1(\rho\eta)\,d\eta, \tag{3·6·1}$$

$$u_z = -\frac{4p_0 c(1-\sigma^2)}{\pi E}\int_0^\infty \left(1+\frac{\zeta\eta}{2(1-\sigma)}\right)\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)e^{-\zeta\eta}J_0(\rho\eta)\,d\eta, \tag{3·6·2}$$

where $\rho = r/c$ and $\zeta = z/c$. Similarly, from equations (3·1·5)–(3·1·8) one obtains for the components of stress at a general point in the interior of the elastic solid

$$\sigma_z = \frac{2p_0}{\pi}\{C_1^0(\rho,\zeta) - S_0^0(\rho,\zeta) + \zeta C_2^0(\rho,\zeta) - \zeta S_1^0(\rho,\zeta)\}, \tag{3·6·3}$$

$$\tau_{zr} = \frac{2p_0\zeta}{\pi}\{C_2^1(\rho,\zeta) - S_1^1(\rho,\zeta)\}, \tag{3·6·4}$$

$$\sigma_r+\sigma_\theta+\sigma_z = \frac{4(1+\sigma)}{\pi}p_0\{C_1^0(\rho,\zeta) - S_0^0(\rho,\zeta)\}, \tag{3·6·5}$$

where C_n^m and S_n^m denote the integrals

$$C_n^m(\rho,\zeta) = \int_0^\infty \eta^{n-1}e^{-\zeta\eta}J_m(\rho\eta)\cos\eta\,d\eta, \quad S_n^m(\rho,\zeta) = \int_0^\infty \eta^{n-1}e^{-\zeta\eta}J_m(\rho\eta)\sin\eta\,d\eta. \tag{3·6·6}$$

The fourth relation for the complete determination of all the components of stress is found from equations (3·1·5) and (3·1·7) to be

$$\sigma_\theta-\sigma_r = 2(\lambda+\mu)\int_0^\infty \xi^3 G'\left\{\frac{2J_1(\xi r)}{\xi r} - J_0(\xi r)\right\}d\xi.$$

Putting $n = 1$, $z = \xi r$ in the recurrence formula

$$J_{n-1}(z) + J_{n+1} = \frac{2n}{z}J_n(z), \tag{3·6·7}$$

then $$\sigma_\theta-\sigma_r = 2(\lambda+\mu)\int_0^\infty \xi^3 G' J_2(\xi r)\,d\xi, \tag{3·6·8}$$

whence follows the relation

$$\sigma_\theta-\sigma_r = \frac{2p_0}{\pi}[(1-2\sigma)\{C_1^2(\rho,\zeta) - S_0^2(\rho,\zeta)\} - \zeta\{C_2^2(\rho,\zeta) - S_1^2(\rho,\zeta)\}]. \tag{3·6·9}$$

Along the axis of symmetry, $r = 0$, it is found that $\tau_{rz} = 0$, $\sigma_r - \sigma_\theta = 0$, so that the maximum shearing stress across any plane through a point on the axis of symmetry is given by the expression $|\tau|$, where

$$\tau = \tfrac{1}{2}(\sigma_z - \sigma_r) = -\frac{p_0}{2\pi}\left\{(1-2\sigma)\left(\int_0^\infty \cos\eta\, e^{-\zeta\eta}\,d\eta - \int_0^\infty \frac{\sin\eta}{\eta} e^{-\zeta\eta}\,d\eta\right)\right.$$
$$\left. + 3\zeta\left(\int_0^\infty \eta\cos\eta\, e^{-\zeta\eta}\,d\eta - \int_0^\infty \sin\eta\, e^{-\zeta\eta}\,d\eta\right)\right\}.$$

Evaluating the integrals

$$\tau = -\tfrac{1}{4}(1-2\sigma)p_0 + \frac{p_0}{2\pi}\left[(1-2\sigma)\left(\frac{\zeta}{1+\zeta^2} + \tan^{-1}\zeta\right) - \frac{6\zeta}{(1+\zeta^2)^2}\right]. \qquad (3{\cdot}6{\cdot}10)$$

Differentiating this expression with respect to ζ then

$$\frac{d\tau}{d\zeta} = \frac{4p_0}{\pi}\frac{(5-\sigma)\zeta^2 - (1+\sigma)}{(1+\zeta^2)^3},$$

showing that $|\tau|$ increases from the value $\tfrac{1}{4}(1-2\sigma)p_0$ at $\zeta = 0$ to a maximum at $\zeta = \{(1+\sigma)/(5-\sigma)\}^{\frac{1}{2}}$ and then decreases steadily to zero at $\zeta = \infty$.

Writing, in the general case, $w = \zeta + i$, then

$$Z_n^m = C_n^m - iS_n^m = \int_0^\infty \eta^{n-1} e^{-w\eta} J_m(\rho\eta)\,d\eta,$$

and the integrals Z_n^m can be evaluated by means of the formulae

$$\int_0^\infty \eta^{n-1} e^{-w\eta} J_m(\rho\eta)\,d\eta = \begin{cases} \dfrac{(n-m+1)!}{(w^2+\rho^2)^{\frac{1}{2}n}} P_{n-1}^m\left(\dfrac{w}{\sqrt{(w^2+\rho^2)}}\right) & (m \leqslant n-1), \\[2ex] \dfrac{(n+m-1)!}{(w^2+\rho^2)^{\frac{1}{2}n}} P_{n-1}^{-m}\left(\dfrac{w}{\sqrt{(w^2+\rho^2)}}\right) & (m > n-1), \end{cases}$$

where P_n^m denotes the associated Legendre function.

In this way one obtains the results

$$Z_1^0 = (\rho^2+w^2)^{-\frac{1}{2}}, \qquad Z_2^0 = w(\rho^2+w^2)^{-\frac{3}{2}},$$
$$Z_0^1 = \frac{1}{\rho}\{(\rho^2+w^2)^{\frac{1}{2}} - w\}, \quad Z_1^1 = \frac{1}{\rho} - \frac{W}{\rho\sqrt{(w^2+\rho^2)}},$$
$$Z_2^1 = \rho(\rho^2+w^2)^{-\frac{3}{2}},$$

and by virtue of equation (3·6·7) the further relation

$$Z_n^2 = \frac{2}{\rho} Z_{n-1}^1 - Z_n^0. \qquad (3{\cdot}6{\cdot}11)$$

Adopting the notation

$$\left.\begin{aligned} r^2 &= 1+\zeta^2, & \zeta\tan\theta &= 1, \\ R^2 &= (\rho^2+\zeta^2-1)^2 + 4\zeta^2, & 2\zeta\cot\phi &= \rho^2+\zeta^2-1, \end{aligned}\right\} \qquad (3{\cdot}6{\cdot}12)$$

and equating real and imaginary parts one obtains the formulae

$$\left.\begin{aligned}
C_0^1 &= R^{-\frac12}\cos\tfrac12\phi, & S_1^0 &= R^{-\frac12}\sin\tfrac12\phi,\\
C_2^0 &= rR^{-\frac32}\cos(\tfrac32\phi-\theta), & S_2^0 &= rR^{-\frac32}\sin(\tfrac32\phi-\theta),\\
C_0^1 &= \frac{1}{\rho}(R^{\frac12}\cos\tfrac12\phi-\zeta), & S_0^1 &= \frac{1}{\rho}(1-R^{\frac12}\sin\tfrac12\phi),\\
C_1^1 &= \frac{1}{\rho}-\frac{r}{\rho}R^{-\frac12}\cos(\theta-\tfrac12\phi), & S_1^1 &= \frac{r}{\rho}R^{-\frac12}\sin(\theta-\tfrac12\phi),\\
C_2^1 &= \rho R^{-\frac32}\cos\tfrac32\phi, & S_2^1 &= \rho R^{-\frac32}\sin\tfrac32\phi,
\end{aligned}\right\} \tag{3·6·13}$$

and from (3·6·12)

$$\left.\begin{aligned}
C_1^2 &= \frac{2}{\rho}C_0^1-C_1^0, \quad S_1^2 = \frac{2}{\rho}S_0^1-S_1^0,\\
C_2^2 &= \frac{2}{\rho}C_1^1-C_2^0.
\end{aligned}\right\} \tag{3·6·14}$$

It only remains to evaluate S_0^0 and S_0^2.

Integrating the expressions for Z_1^0, Z_0^1 with respect to w, then

$$\int_0^\infty \frac{1-e^{-wx}}{x}J_0(\rho x)\,dx = \log\frac{\sqrt{(\rho^2+w^2)}+w}{\rho}, \tag{3·6·15}$$

$$\int_0^\infty \frac{1-e^{-wx}}{x^2}J_1(\rho x)\,dx = \frac{1}{2\rho}\left[w\sqrt{(w^2+\rho^2)}-w^2+\rho^2\log\frac{\sqrt{(\rho^2+w^2)}+w}{\rho}\right], \tag{3·6·16}$$

it being assumed in both cases that $\rho \neq 0$. With the notation of (3·6·12) then from (3·6·16) by equating imaginary parts

$$S_0^0 = \tan^{-1}\frac{R^{\frac12}\sin\frac12\phi+r\sin\theta}{R^{\frac12}\cos\frac12\phi+r\cos\theta} \quad (\rho\neq 0). \tag{3·6·17}$$

Using the recurrence relation (3·6·7) it is deduced, from equations (3·6·15) and (3·6·16), that if $\rho \neq 0$,

$$\int_0^\infty \frac{1-e^{-wx}}{x}J_2(\rho x)\,dx = \frac{1}{\rho^2}\{w\sqrt{(w^2+\rho^2)}-w^2\} = wZ_0^1/\rho.$$

Putting $w = \zeta+i$, $Z_0^1 = C_0^1-iS_0^1$ and equating imaginary parts, then

$$S_0^2 = \frac{1}{\rho}(C_0^1-\zeta S_0^1). \tag{3·6·18}$$

3·7. Before proceeding to the determination of the distribution of stress in the medium at points remote from the crack consider the expressions for the stress components at points in the immediate vicinity of the periphery of the crack. In this way results are obtained analogous to those of § 1·4; with a notation similar to that of that section the integrals of § 3·6 are computed when $\zeta = \delta \sin \psi / c$, and $\rho = 1 + \delta \cos \psi / c$. Then

$$\left.\begin{aligned} r &= 1, & \phi &= \psi - \tfrac{1}{2}\delta \sin \psi / c, \\ \theta &= \tfrac{1}{2}\pi - \delta \sin \psi / c, & R &= 2\delta(c + \delta \cos \psi)/c^2. \end{aligned}\right\} \qquad (3\cdot 7\cdot 1)$$

Substituting from equations (3·7·1) into the formulae (3·6·13–18) and retaining only those terms which are of first or higher order in $(c/\delta)^{\frac{1}{2}}$, it follows that

$$S_1^1 = C_1^0 = -C_1^2 = \left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \cos \tfrac{1}{2}\psi, \quad C_2^1 = \left(\frac{c}{2\delta}\right)^{\frac{3}{2}} \cos \tfrac{3}{2}\psi,$$

$$C_2^0 = -C_2^2 = \left(\frac{c}{2\delta}\right)^{\frac{3}{2}} \sin \tfrac{3}{2}\psi, \qquad S_1^0 = -S_1^2 = \left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \sin \tfrac{1}{2}\psi,$$

the other integrals being negligible to this approximation. With these values for the integrals involved the following expressions are obtained from equations (3·6·3–5) and (3·6·9) for the stress components σ_r, σ_z and τ_{rz}:

$$\sigma_r = \frac{2p_0}{\pi}\left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \{\tfrac{3}{4} \cos \tfrac{1}{2}\psi + \tfrac{1}{4} \cos \tfrac{5}{2}\psi\}, \qquad (3\cdot 7\cdot 2)$$

$$\sigma_z = \frac{2p_0}{\pi}\left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \{\tfrac{5}{4} \cos \tfrac{1}{2}\psi - \tfrac{1}{4} \cos \tfrac{5}{2}\psi\}, \qquad (3\cdot 7\cdot 3)$$

$$\tau_{zr} = \frac{p_0}{\pi}\left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \sin \psi \cos \tfrac{3}{2}\psi. \qquad (3\cdot 7\cdot 4)$$

It will be observed that these formulae differ from the expressions for the components of stress in the two-dimensional case (equations (1·4·1–3)) only by the presence of a factor $2/\pi$. There is no parallel in the two-dimensional analysis to the hoop stress σ_θ; in the three-dimensional case its value near the edge of the crack is given by the equation

$$\sigma_\theta = \frac{4\sigma p_0}{\pi}\left(\frac{c}{2\delta}\right)^{\frac{1}{2}} \cos \tfrac{1}{2}\psi. \qquad (3\cdot 7\cdot 5)$$

3·8. By means of the formulae of § 3·6 the components of stress at any point $r = \rho a$, $z = \zeta a$, in the interior of the elastic body can be calculated for any prescribed value of the Poisson ratio. The results of these calculations for the case where the Poisson ratio has the value 0·25 and where both ρ and ζ assume a sequence of values are given in tables 2–5. The variation of the various components of stress in certain planes is shown graphically in figures 5–8. The rapid decay of the stress components σ_r and σ_θ as z increases is illustrated by figures 5 and 6; the variation with z of the normal component of stress, σ_z, is much more gradual as is seen from figure 7. Perhaps the most striking feature of these results is the marked similarity between the variation of the shearing stress τ_{zr} and that of the principal shearing stress τ in the two-dimensional case (compare figures 1 and 8).

Table 2. The variation of σ_r/p_0 with ρ and ζ

ζ \ ρ	0·0	0·2	0·4	0·6	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·0	−0·750	−0·750	−0·750	−0·750	−0·750	∞	0·010	−0·055	−0·068	−0·069	−0·066
0·2	−0·446	−0·433	−0·386	−0·297	−0·180	−0·070	0·007	0·060	0·072	0·054	0·039
0·4	−0·214	−0·207	−0·160	−0·121	−0·087	−0·057	−0·041	−0·019	0·000	0·018	0·014
0·6	−0·075	−0·068	−0·060	−0·050	−0·039	−0·029	−0·019	−0·013	−0·007	0·000	0·007
0·8	−0·006	−0·001	0·007	0·013	0·022	0·014	0·006	0·001	0·000	−0·016	−0·026
1·0	0·017	0·015	0·012	0·010	0·005	−0·011	−0·022	−0·026	−0·028	−0·031	−0·027
1·2	0·031	0·029	0·026	0·023	0·022	0·019	0·015	0·009	−0·001	−0·007	−0·009
1·4	0·030	0·028	0·025	0·020	0·015	0·009	0·005	0·002	0·000	−0·001	−0·001
1·6	0·028	0·027	0·025	0·020	0·015	0·011	0·007	0·003	0·001	0·000	−0·001
1·8	0·024	0·023	0·022	0·019	0·015	0·010	0·007	0·004	0·001	0·001	0·000
2·0	0·021	0·020	0·018	0·016	0·013	0·010	0·006	0·005	0·003	0·002	0·000
3·0	0·009	0·009	0·008	0·008	0·007	0·006	0·005	0·004	0·003	0·002	0·002
4·0	0·004	0·004	0·004	0·004	0·003	0·003	0·003	0·003	0·002	0·002	0·002

Table 3. The variation of σ_θ/p_0 with ρ and ζ

ζ \ ρ	0·0	0·2	0·4	0·6	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·0	−0·750	−0·750	−0·750	−0·750	−0·750	∞	0·010	−0·055	−0·068	−0·069	−0·066
0·2	−0·446	−0·438	−0·412	−0·344	−0·242	−0·100	0·045	0·022	0·010	0·005	0·004
0·4	−0·214	−0·219	−0·178	−0·131	−0·067	−0·014	0·025	0·018	0·013	0·009	0·006
0·6	−0·075	−0·072	−0·051	−0·009	0·016	0·020	0·023	0·013	0·007	0·006	0·005
0·8	−0·006	−0·009	−0·017	−0·023	−0·019	−0·008	0·009	0·011	0·014	0·017	0·012
1·0	0·017	0·019	0·026	0·033	0·030	0·026	0·022	0·018	0·017	0·015	0·003
1·2	0·031	0·039	0·045	0·053	0·061	0·058	0·039	0·021	0·011	0·019	0·006
1·4	0·030	0·035	0·039	0·044	0·048	0·054	0·048	0·041	0·033	0·025	0·016
1·6	0·028	0·032	0·037	0·040	0·043	0·046	0·042	0·038	0·033	0·029	0·021
1·8	0·024	0·027	0·029	0·032	0·034	0·036	0·038	0·036	0·032	0·028	0·020
2·0	0·021	0·023	0·025	0·027	0·029	0·031	0·032	0·030	0·027	0·025	0·019
3·0	0·009	0·009	0·010	0·010	0·011	0·011	0·013	0·012	0·011	0·010	0·010
4·0	0·004	0·005	0·005	0·006	0·006	0·006	0·006	0·006	0·006	0·005	0·005

Table 4. The variation of σ_z/p_0 with ζ and ρ

ζ \ ρ	0·0	0·2	0·4	0·6	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·0	−1·000	−1·000	−1·000	−1·000	−1·000	∞	0·333	0·143	0·080	0·050	0·034
0·2	−0·987	−0·985	−0·975	−0·937	−0·729	0·232	0·358	0·166	0·089	0·055	0·035
0·4	−0·917	−0·912	−0·862	−0·742	−0·451	−0·006	0·189	0·150	0·095	0·060	0·040
0·6	−0·788	−0·768	−0·700	−0·560	−0·334	−0·081	0·068	0·097	0·080	0·053	0·041
0·8	−0·639	−0·617	−0·548	−0·429	−0·263	−0·108	0·002	0·048	0·055	0·048	0·038
1·0	−0·508	−0·482	−0·424	−0·333	−0·237	−0·113	−0·032	0·014	0·018	0·022	0·030
1·2	−0·386	−0·373	−0·329	−0·262	−0·185	−0·109	−0·049	−0·009	0·012	0·021	0·023
1·4	−0·302	−0·286	−0·256	−0·210	−0·154	−0·094	−0·055	−0·022	−0·001	0·010	0·015
1·6	−0·230	−0·224	−0·201	−0·168	−0·129	−0·090	−0·055	−0·027	−0·010	0·001	0·008
1·8	−0·180	−0·175	−0·159	−0·138	−0·108	−0·079	−0·054	−0·032	−0·010	−0·005	0·002
2·0	−0·142	−0·139	−0·127	−0·111	−0·091	−0·070	−0·050	−0·033	−0·019	−0·012	−0·003
3·0	−0·052	−0·051	−0·049	−0·045	−0·040	−0·035	−0·029	−0·024	−0·019	−0·014	−0·010
4·0	−0·025	−0·024	−0·023	−0·022	−0·021	−0·019	−0·017	−0·014	−0·013	−0·011	−0·009

TABLE 5. THE VARIATION OF τ_{zr}/p_0 WITH ρ AND ζ

ζ \ ρ	0·0	0·2	0·4	0·6	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·0	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000
0·2	0·000	−0·020	−0·051	−0·121	−0·327	−0·420	0·066	0·052	0·030	0·017	0·011
0·4	0·000	−0·055	−0·126	−0·233	−0·354	−0·319	−0·097	0·003	0·021	0·019	0·014
0·6	0·000	−0·074	−0·155	−0·239	−0·292	−0·252	−0·133	−0·041	−0·006	0·018	0·007
0·8	0·000	−0·074	−0·145	−0·202	−0·225	−0·200	−0·131	−0·069	−0·029	−0·009	0·000
1·0	0·000	−0·061	−0·119	−0·159	−0·170	−0·156	−0·117	−0·075	−0·046	−0·021	−0·009
1·2	0·000	−0·050	−0·092	−0·123	−0·133	−0·124	−0·100	−0·076	−0·047	−0·028	−0·016
1·4	0·000	−0·040	−0·070	−0·092	−0·102	−0·097	−0·084	−0·065	−0·048	−0·032	−0·020
1·6	0·000	−0·029	−0·053	−0·070	−0·078	−0·077	−0·069	−0·057	−0·044	−0·032	−0·022
1·8	0·000	−0·021	−0·040	−0·053	−0·060	−0·061	−0·057	−0·049	−0·040	−0·031	−0·023
2·0	0·000	−0·016	−0·031	−0·041	−0·047	−0·049	−0·046	−0·042	−0·035	−0·029	−0·022
3·0	0·000	−0·005	−0·009	−0·012	−0·015	−0·017	−0·018	−0·018	−0·017	−0·016	−0·014
4·0	0·000	−0·002	−0·004	−0·005	−0·006	−0·007	−0·007	−0·008	−0·008	−0·008	−0·008

The principal stresses at any point are determined by the roots of the discriminating cubic

$$\begin{vmatrix} \sigma-\sigma_r & 0 & \tau_{zr} \\ 0 & \sigma-\sigma_\theta & 0 \\ \tau_{zr} & 0 & \sigma-\sigma_z \end{vmatrix} = 0, \qquad (3{\cdot}8{\cdot}1)$$

so that they are equal to

$$\sigma_\theta, \quad \tfrac{1}{2}(\sigma_r+\sigma_z)\pm\{(\tfrac{1}{2}\sigma_r-\tfrac{1}{2}\sigma_z)^2+\tau_{rz}^2\}^{\frac{1}{2}}. \qquad (3{\cdot}8{\cdot}2)$$

With the values of the stress components given in tables 2–5 the components of principal stress may readily be deduced from the formulae (3·8·2). The numerical value of the principal shearing stress can then be determined from the fact that it is equal to one-half of the algebraic difference between the maximum and the minimum components of the principal stress. The results of such a calculation are embodied in table 6, and the variation of the principal shearing stress in planes parallel to the plane of the crack exhibited graphically in figure 9. In general outline this variation is similar to that of the principal shearing stress in the two-dimensional case.

TABLE 6. THE VARIATION OF THE PRINCIPAL SHEARING STRESS (DIVIDED BY p_0) WITH ρ AND ζ

ζ \ ρ	0·0	0·2	0·4	0·6	0·8	1·0	1·2	1·4	1·6	1·8	2·0
0·0	0·125	0·125	0·125	0·125	0·125	∞	0·172	0·099	0·074	0·059	0·050
0·2	0·270	0·277	0·299	0·342	0·427	0·446	0·188	0·085	0·051	0·033	0·021
0·4	0·351	0·356	0·373	0·388	0·398	0·320	0·151	0·085	0·047	0·033	0·023
0·6	0·357	0·358	0·356	0·349	0·327	0·254	0·140	0·069	0·044	0·032	0·025
0·8	0·317	0·316	0·313	0·299	0·266	0·201	0·131	0·073	0·040	0·032	0·032
1·0	0·263	0·256	0·248	0·234	0·209	0·164	0·117	0·075	0·044	0·031	0·009
1·2	0·209	0·207	0·200	0·188	0·168	0·139	0·105	0·077	0·048	0·030	0·017
1·4	0·169	0·162	0·157	0·147	0·123	0·110	0·089	0·066	0·048	0·032	0·022
1·6	0·129	0·129	0·125	0·117	0·107	0·092	0·076	0·059	0·045	0·032	0·023
1·8	0·102	0·100	0·099	0·095	0·086	0·076	0·064	0·052	0·040	0·031	0·023
2·0	0·082	0·081	0·079	0·075	0·070	0·063	0·054	0·046	0·037	0·030	0·022
3·0	0·030	0·030	0·030	0·029	0·028	0·026	0·025	0·023	0·020	0·018	0·016

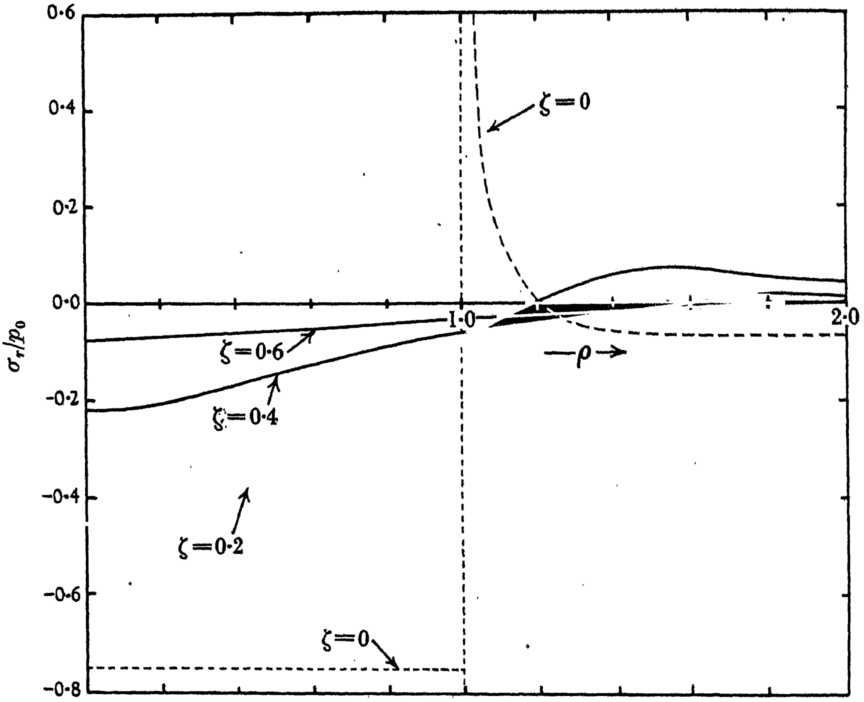


FIGURE 5. The variation of the radial component of stress, σ_r, with ρ and ζ. The broken curve shows the variation of σ_r in the plane of the crack ($z = 0$).

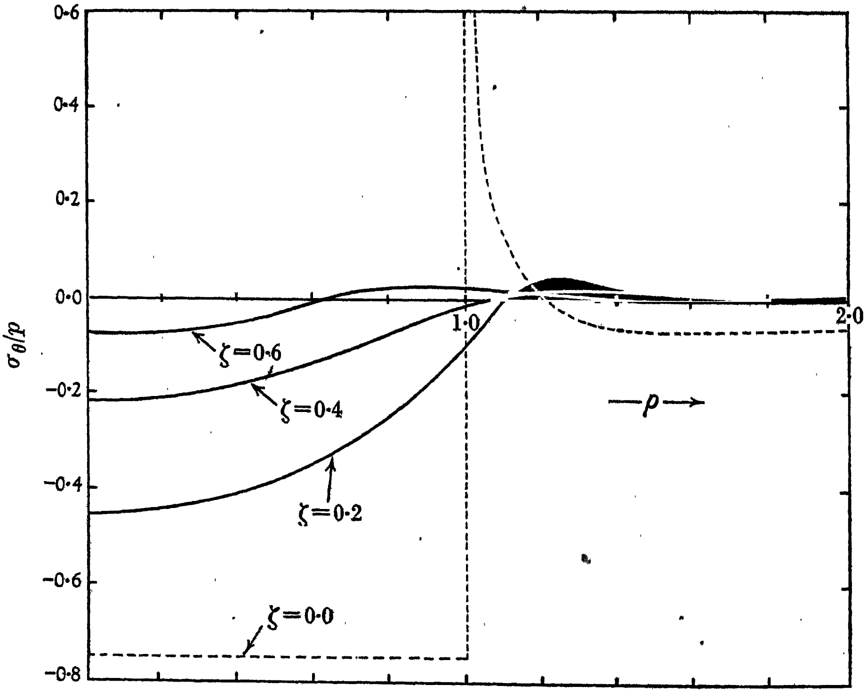


FIGURE 6. The variation of the hoop stress, σ_θ, with ρ and ζ. The broken curve shows the variation of σ_θ in the plane of the crack ($z = 0$).

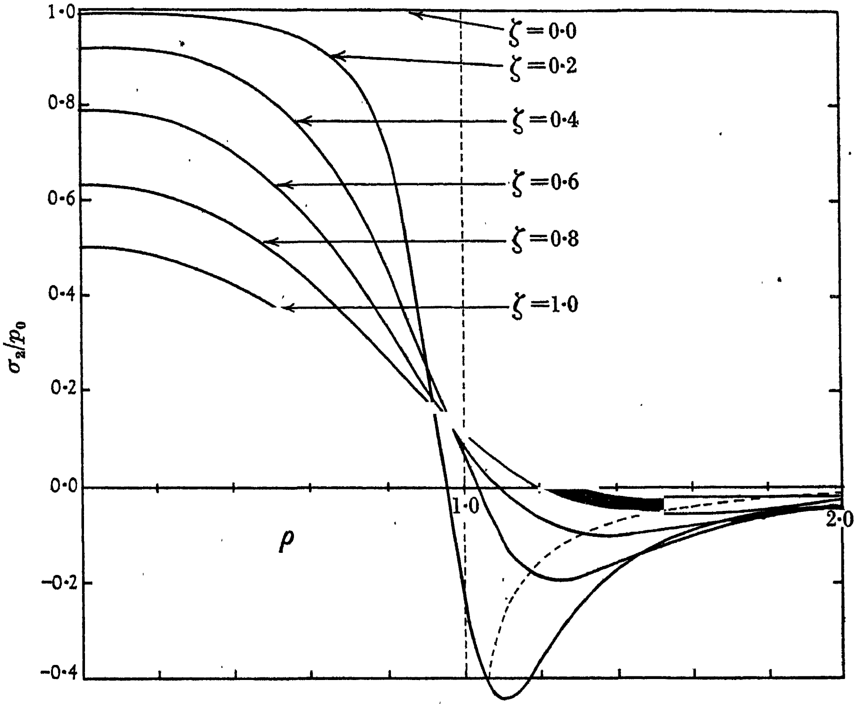


FIGURE 7. The variation of the normal component of stress, σ_z, with ρ and ζ. The broken curve shows the variation of σ_z in the plane of the crack ($z = 0$).

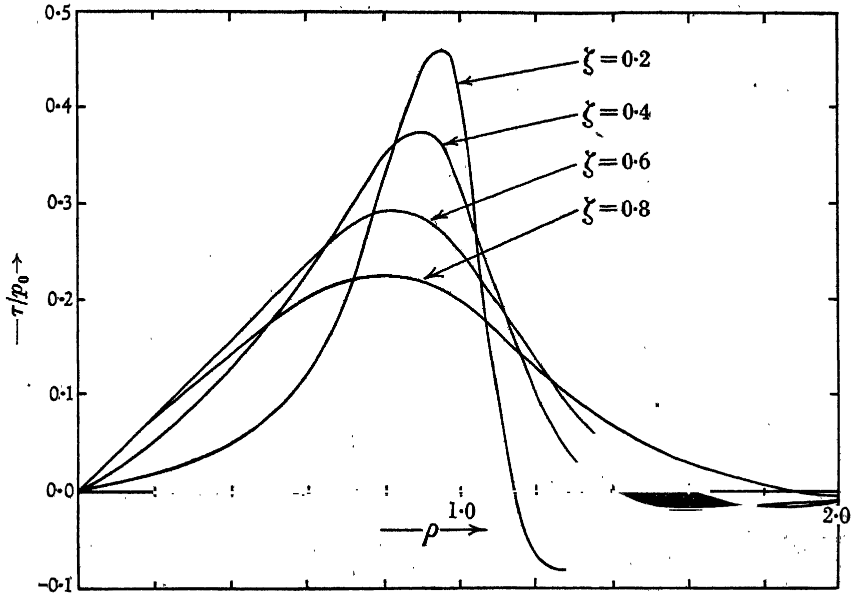


FIGURE 8. The variation of the shearing stress, τ_{zr}, with ρ and ζ.

The most convenient way of visualizing the distribution of stress in an elastic body is probably to draw the contours of equal principal shearing stress—the three-dimensional analogue of the isochromatic lines of photoelasticity. By means of table 6 and figure 9, points in the z-r plane at which the principal shearing stress

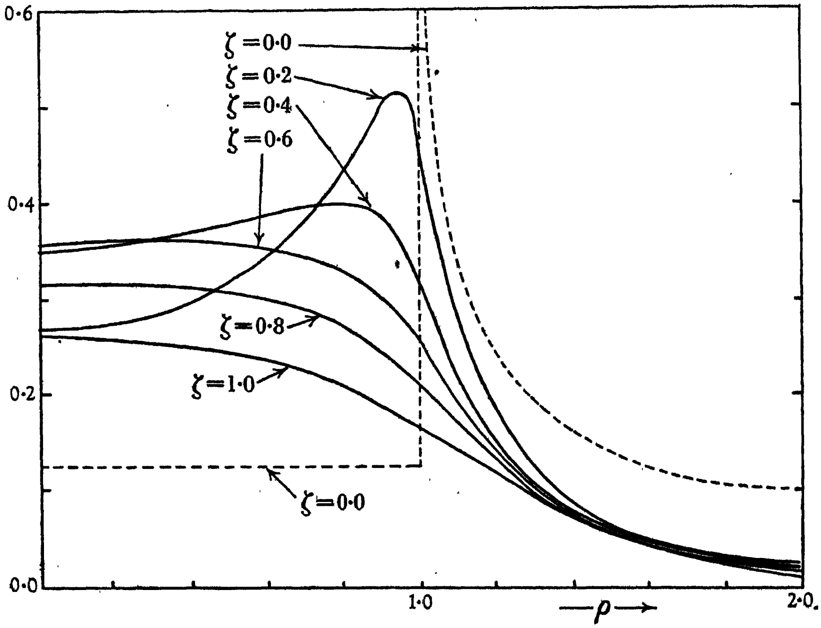


FIGURE 9. The variation of the principal shearing stress, τ, with ρ and ζ. The broken curve shows the variation of τ in the plane of the crack.

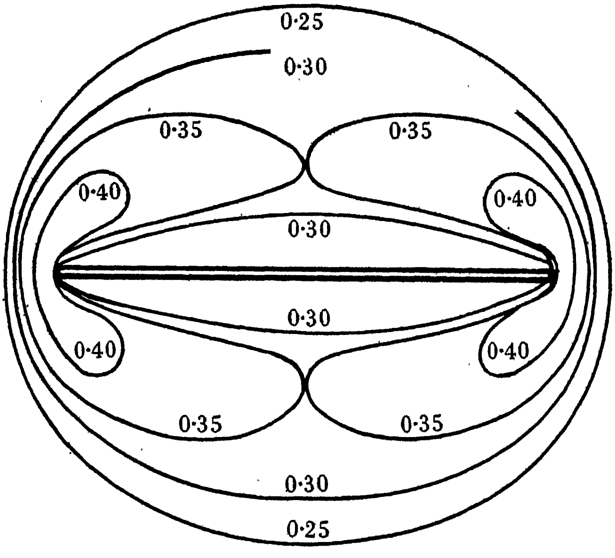


FIGURE 10. The contours of equal principal shearing stress in the vicinity of a circular crack.

reaches the values 0·25, 0·30, 0·35, 0·40 can readily be determined (in terms of p_0 as unit), and by drawing the curves through corresponding values the contours of (equal) principal shearing stress can be plotted. The result is shown in figure 10. It will be observed that except in the vicinity of the edges of the crack—and apart from numerical values—these contours of equal principal shearing stress are very like the isochromatic lines of the two-dimensional case (figure 2). The main difference between the two systems of curves lies in their behaviour near the edges of the crack. In the two-dimensional case the points $x = \pm c$ are nodes separating two loops of the curves (cf. the curves corresponding to the values $\tau/p_0 = 0{\cdot}4, 0{\cdot}5, 0{\cdot}7$ in figure 2), but in the three-dimensional case the point $r = c$ is a simple point of the curve, the curve cutting the r-axis again at a point where $r > c$ (cf. curves with τ/p_0 in figure 10). It should, of course, be remembered that these 'contours' are in reality *surfaces* of equal principal shearing stress obtained by rotating figure 10 about the z-axis.

As in the two-dimensional case the principal shearing stress becomes infinite at the edges of the crack, indicating that a certain amount of plastic flow will occur and hence that there is in reality no purely elastic problem. Except in the neighbourhood of the crack, however, the elastic stresses are dominant and the energy relation (3·3·6) approximates very closely to the true result.

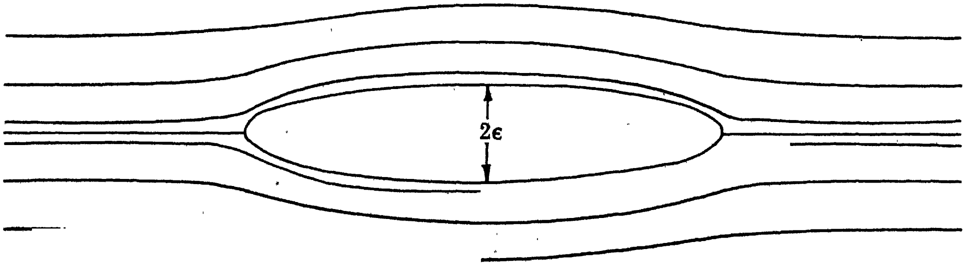


FIGURE 11. The normal displacement of planes parallel to the plane of an ellipsoidal crack of small depth. The depth of the crack, 2ϵ, has been greatly exaggerated in this diagram to show the variation of the normal displacement more clearly.

3·9. The distribution of stress in the vicinity of a crack can also be illustrated by calculating the normal component of the displacement vector for points in the elastic body and then drawing the normal displacement of planes which were originally parallel to the plane of the crack. Since the normal component of stress σ_z is infinitely large at the edge of the crack it might at first sight appear that the displacement u_z should also be infinite in that region; that this is not so can readily be seen from an examination of equation (3·6·2). It follows immediately from this equation that the normal component of the displacement vector is given by the formula

$$u_z = \epsilon\left[1-\rho S_0^1-\zeta S_0^0-\frac{\zeta}{2(1-\sigma)}(\zeta S_1^0+\rho S_1^1-S_0^0)\right] \qquad (3{\cdot}9{\cdot}1)$$

with the notation of (3·6·6). By means of this formula the variation of the normal displacement along the planes $z/c = 0{\cdot}05, 0{\cdot}2, 0{\cdot}4$ was calculated. The results are shown graphically in figure 11, which shows the position of the planes $z/c = 0{\cdot}05, 0{\cdot}2,$

0·4 after the crack has been opened out to a depth 2ϵ, by a uniform internal pressure; in this diagram the ratio ϵ/c has been very much exaggerated to show more clearly the variation of the normal displacement in planes near to the crack.

Approximate formulae for the determination of the normal component of the displacement vector in planes close to the crack can be derived from equation (3·9·1). If ρ is small, then

$$J_0(\rho\eta) = 1-\tfrac{1}{4}\rho^2\eta^2, \quad J_1(\rho\eta) = \tfrac{1}{2}\rho\eta$$

may be written in the integrals for C_n^m and S_n^m; making this substitution, evaluating the integrals now involved and retaining terms of (3·9·1) of order two or less one obtains the approximate result

$$u_z = \epsilon\{f_1(\zeta)-\rho^2 f_2(\zeta)\}+O(\epsilon\rho^4), \tag{3·9·2}$$

where the functions $f_1(\zeta)$ and $f_2(\zeta)$ are defined by the relations

$$\left.\begin{aligned} f_1(\zeta) &= 1-\frac{(1-2\sigma)}{2(1-\sigma)}\zeta\left(\frac{\pi}{2}-\tan^{-1}\zeta\right)-\frac{1}{2(1-\sigma)}\frac{\zeta^2}{1+\zeta^2},\\ f_2(\zeta) &= \left(\frac{1}{2}-\frac{3-\sigma}{2(1-\sigma)}\zeta^2\right)(1+\zeta^2)^{-3}. \end{aligned}\right\} \tag{3·9·3}$$

By means of equation (3·9·2) may be calculated the normal displacement at points just off the axis of symmetry $r = 0$.

Similarly, by taking ρ to be large the integrals involved by means of equations (3·6·13) can be evaluated; since as a first approximation one may take $R=\rho$, $\phi = 2\zeta/\rho^2$ in these equations. Thus for ρ large and $\zeta \ll \rho$ one has the approximate formula

$$u_z = \frac{\epsilon\zeta}{\rho}\left\{1-\frac{\zeta}{2(1-\sigma)}\right\}+O\left(\frac{\epsilon}{\rho^2}\right). \tag{3·9·4}$$

The value of the displacement vector in the neighbourhood of the periphery of the crack $r=c$ can be determined by the methods of § 3·7. With the notation of that section

$$u_z = \epsilon\sqrt{\left(\frac{2\delta}{c}\right)}\sin\tfrac{1}{2}\psi\left[1-\frac{1+\cos\psi}{4(1-\sigma)}\right], \tag{3·9·5}$$

where $z = \delta\sin\psi$ and $r = c+\delta\cos\psi$.

The relation of the approximate formulae (3·9·2–5) to the exact result (3·9·1) is shown in figure 12. In this diagram the value of the displacement u_z for points along the plane $z/c = 0·05$ are shown; the full line gives the values of the displacement as calculated by equation (3·9·1) for $\sigma = 0·25$; the broken lines show the values given by the approximate formulae, for the same value of the Poisson ratio. It will be observed that the agreement is good within the range specified. It is to be expected that for smaller values of δ the expression (3·9·5) will give more accurate results.

3·10. The effect of a circular crack on the strength of a brittle solid subjected to a uniform tensile stress can be deduced immediately from the analysis of §§ 3·5 and 3·6. Assume that the radius c of the crack (which may be taken to be situated in the plane $z = 0$ with its centre at the origin of co-ordinates) is considerably smaller than the dimensions of the elastic body. Then replace the finite elastic body by an infinite elastic body with a crack in its interior which is subjected to the stress system

$$\sigma_z = p_0, \quad \sigma_r = \sigma_\theta = p_1 \tag{3·10·1}$$

when r is very large. Since the surface of the crack may be assumed to be free from external forces the further boundary conditions

$$\sigma_z = 0, \quad \tau_{zr} = 0 \tag{3·10·2}$$

follow when $z = 0$ and $r < c$.

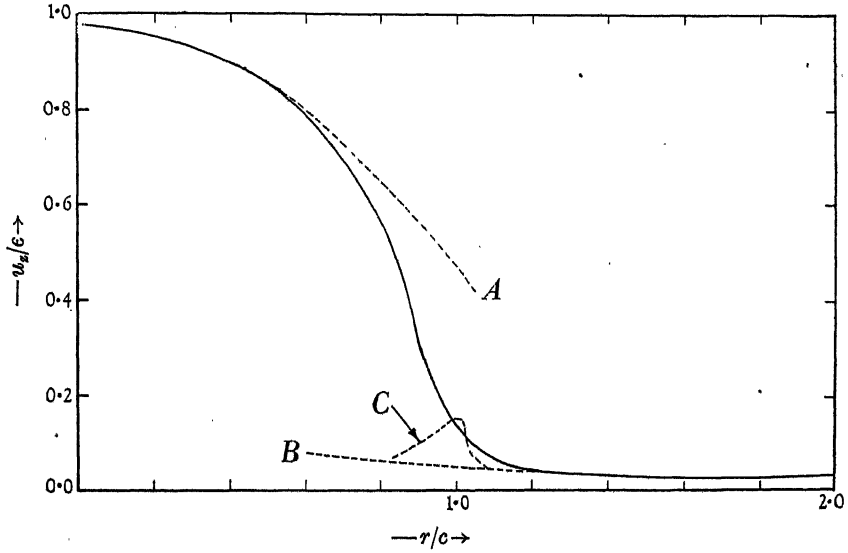


FIGURE 12. The variation of the normal displacement u_z in the plane $z = 0·05$ c.c. The full curve corresponds to the exact values calculated from equation (3·9·1). The broken curves correspond to the approximate formulae—A to (3·9·2), B to (3·9·4) and C to (3·9·5).

If the solution
$$\tau_{zr} = 0, \quad \sigma_z = p_0, \quad \sigma_r = \sigma_\theta = p_1 \tag{3·10·3}$$

be added to equations (3·6·3), (3·6·4), (3·6·5) and (3·6·9), the expressions so obtained for the components of stress satisfy the equations of equilibrium (3·1·3) and (3·1·4) and the boundary conditions (3·10·1) and (3·10·2). The distribution of stress in the interior of the solid can then be determined from the results of § 3·7 by a very simple calculation, or from figures 5–8 by a simple change of scale along the axis corresponding to the stress component.

It follows immediately from equation (3·3·6) that if a brittle body is acted upon by uniform stress $\sigma_r = \sigma_\theta = p_1$, $\sigma_z = p_0$, the presence of a circular crack of radius c lowers the potential energy of the medium by an amount

$$W = \frac{8(1-\sigma^2)p_0^2c^3}{3E} \tag{3·10·4}$$

which is independent of the value of p_1. The crack has also a surface energy

$$U = 2\pi c^2 T, \tag{3·10·5}$$

where T denotes the surface tension of the material of the elastic body. The condition that the crack may spread is

$$\frac{\partial}{\partial c}(W-U) = 0$$

or

$$c = \frac{\pi E T}{2p_0^2(1-\sigma^2)},$$

so that the crack will become unstable and spread if p_0 exceeds the critical value

$$p_{cr} = \sqrt{\frac{\pi E T}{2c(1-\sigma^2)}}.$$

This is the criterion derived by Sack (1946) (equation (1·2·1) above).

I should like to take this opportunity of expressing my thanks to Professor N. F. Mott for acquainting me with this problem and for various suggestions received during many helpful discussions on the theory of cracks. I am also indebted to Mr R. Sack, of the H. H. Wills Physical Laboratory, the University of Bristol, for putting his unpublished work at my disposal, and to Miss June Tenwick for her assistance in checking the calculations of § 3·8.

REFERENCES

Busbridge, I. W. 1938 *Proc. Lond. Math. Soc.* (2), **44**, 115.
Frocht, M. M. 1941 *Photoelasticity*, p. 347. New York: McGraw Hill.
Griffith, A. A. 1920 *Phil. Trans.* A, **221**, 163.
Griffith, A. A. 1924 *Proc. Int. Congr. Appl. Mech. Delft*, p. 55.
Harding, J. W. & Sneddon, I. N. 1945 *Proc. Camb. Phil. Soc.* **41**, 16.
Inglis, C. E. 1913*a* *Engineering*, **95**, 415.
Inglis, C. E. 1913*b* *Trans. Instn Naval Archit.* **55**, 219.
Jahnke, E. & Emde, F. 1938 *Tables of Functions*. Leipzig: Teubner.
Love, A. E. H. 1934 *Mathematical theory of elasticity*, 4th ed. Cambridge.
Nadai, A. 1931 *Plasticity*. New York: McGraw Hill.
Neuber, H. 1934 *Z. angew. Math. Mech.* **14**, 203.
Orowan, E. 1934 *Zs. Kristallogr.* **89**, 327.
Sack, R. 1946 To be published shortly.
Sneddon, I. N. 1945 *Proc. Camb. Phil. Soc.* **42**, 29.
Titchmarsh, E. C. 1937 *An introduction to the theory of Fourier integrals.* Oxford Univ. Press.
Watson, G. N. 1944 *The theory of Bessel functions*, 2nd ed. Cambridge Univ. Press.
Westergaard, H. M. 1939 *J. Appl. Mech.* **6**, A, 49.

The basic reactions in the upper atmosphere. I

By D. R. Bates, M.Sc. and H. S. W. Massey, F.R.S.

University College, London

(*Received* 7 *November* 1945—*Read* 14 *March* 1946)

The properties of the upper atmosphere and in particular of the ionosphere and of the night-sky emission are outlined. The fundamental processes that might be involved are discussed. Various suggestions regarding the production of the layers are considered in detail and the main difficulties indicated. The theory of the equilibrium of the layers is built up and the outstanding problem of explaining the rate of disappearance of electrons is emphasized. Current views on the origin of the night-sky light are studied critically and serious discrepancies are shown to exist. Other possibilities are examined.

1. In an earlier paper (Massey 1937) the nature of the collision processes occurring in the ionosphere was investigated in general terms with a view to understanding the observed distribution of electron and ion density in space and time. Explicit account was taken of the known stability of the negative ions of oxygen, so that besides direct recombination of electrons and positive ions, consideration was given to attachment of electrons to O and O_2, detachment of electrons from O^- and O_2^-, and recombination of negative and positive ions. In agreement with Martyn & Pulley (1936) it was concluded that negative-ion phenomena are likely to play an important role in the ionosphere. From the preliminary examination of the rates of the various processes as derived from theoretical and experimental data it was suggested that the best way of regarding the situation in the *E* layer is to take the electrons and negative ions as in a dynamical equilibrium so that if n_e, n^- are their respective densities,

$$n^-/n_e = \lambda, \tag{1}$$

where λ changes only slowly with time. The apparent recombination coefficient for electrons is then given by

$$\alpha = \alpha_e + \lambda\alpha_i, \tag{2}$$

where α_e is the true coefficient of recombination between electrons and positive ions and α_i that for recombination between negative and positive ions.

In a second paper (Bates, Buckingham, Massey & Unwin 1939) α_e was calculated by quantal methods and found to be quite inadequate to account for the observed value of α. Rough estimates of α_i (now thought to be too low) then suggested that λ in (2) must be taken of order 100. The difficulty of reconciling this with photo-detachment effects during the daytime was not properly realized. At that time also it appeared that a large value of λ was required to explain the magnitude of the terrestrial magnetic variations, but the measurements of Appleton & Weeks (1939) on atmospheric tides now cast doubt on the validity of the theoretical treatment of this phenomenon.

For further progress it was necessary to extend the application of theoretical analysis to as many reaction rates as possible. As part of such a programme a detailed study of O^- and O_2^- and of the reactions leading to their formation and destruction has recently been made by the present authors (Bates & Massey 1943). From this work and that of the previous paper (Bates *et al.* 1939) information is available concerning the continuous absorption cross-section of atomic oxygen, the electronic and ionic recombination coefficients, the rate of photo-detachment and a number of other processes. With this for guidance it becomes worth while to review the theory of the upper atmosphere taking into account also the recent observational material. The distribution of solar photo-ionization throughout the atmosphere and its relation to the origin of the ionized layers, the temporal variation of ion density, and the light of the night sky are all discussed. Considerable modifications are introduced by such factors as the intense absorption estimated for atomic oxygen, the large value the ionic recombination coefficient may assume, and the previously underestimated importance of photo-detachment of electrons from negative ions during daylight.* A wholly satisfactory theory cannot yet be regarded as available and, in all cases, care has been taken to catalogue and discuss the difficulties still outstanding. Directions in which information should be sought are outlined.

2. MEAN COMPOSITION, TEMPERATURE, AND DENSITY DISTRIBUTION OF THE ATMOSPHERE

2·1. *Composition*

The principal constituents of the atmosphere are oxygen and nitrogen. As pointed out by Chapman (1931), oxygen is mainly present as molecules at low heights and as atoms at great heights. The detailed theory of Wulf & Deming (1938) shows that the transition takes place in a region near but slightly above 100 km.† The existence of oxygen in the atomic state in the high atmosphere has been confirmed by the observation of emission lines in the night-sky and auroral spectra‡ (§ 7·1·1). Nitrogen however, probably remains in the molecular state up to very great heights. There is no reason to expect, nor evidence to suggest, dissociation (Chapman & Price 1936). For lack of knowledge of the rate of mixing, and consequently of the importance of diffusion, it is generally taken as a crude approximation that the relative proportions of oxygen and nitrogen remain constant—that is, the ratio $O_2/N_2 = \frac{1}{4}$

---

* The effect of photo-detachment in reducing the value of λ to well below unity has been reported independently by Massey & Bates (1943) and Ta-You Wu (1945).

† Slight adjustment has here been made to the original results of Wulf & Deming to allow for the fact that the density of their model atmosphere falls off more rapidly with height than is indicated by present evidence (§ 2·3).

‡ It may also be significant that the band systems of molecular oxygen though possibly present in the night-sky spectrum (emitted from an unknown height) are apparently absent from the auroral spectrum (emitted mainly from heights between about 100 and 200 km.) (§ 7·1·5).

below the atmospheric transition region, and the ratio $O/N_2 = \frac{1}{2}$ above the atmospheric transition region. The most promising source of information on the oxygen-nitrogen distribution appears to be from careful observation and thorough understanding of the auroral spectrum.

Various diatomic compounds of oxygen, nitrogen, carbon and hydrogen might be expected to occur in the high atmosphere. Band emissions arising from nitric oxide have been reported in the night-sky spectrum (§ 7·1·4), but there appears little evidence of any from the other possible compounds. Certain polyatomic molecules are present. Carbon dioxide and water vapour are contained in the lower atmosphere, but the height to which they extend is uncertain. The total amount of the former is some 6×10^{21} mol./cm.^{2} column of atmosphere, that of the latter is of course very variable. Ozone is well known to be present in a layer mainly between 20 and 30 km. Absorption measurements show the number of molecules to be of the order 7×10^{18}/cm.^{2} column of atmosphere (Chapman 1943). Nitrous oxide is also to be expected, and there is the evidence for its presence in the absorption spectrum (Sutherland & Callendar 1943). These subsidiary constituents may affect profoundly the temperature distribution, but they have probably not much direct influence on the phenomena discussed in this paper (see, however, § 4·6).

Attention has been drawn in recent years to the presence in the atmosphere of foreign elements originating presumably from meteors or terrestrial dust. The best authenticated of these is sodium, which gives rise to the striking D lines in the night-sky spectrum. From the intensity of the twilight enhancement effect it can be estimated that above about 70 km. the number of sodium atoms/cm.^{2} column of atmosphere is about 5×10^8 to 5×10^9 (§ 7·2·2).

2·2. *Temperature*

The temperature distribution has been fully considered by Martyn & Pulley (1936) and more recently by Whipple (1938). From the evidence obtained from balloon exploration, anomalous sound propagation, noctilucent clouds, the nitrogen bands in the auroral spectrum, and other phenomena, the latter concluded that the temperature rises from about 220° K in the stratosphere to a broad maximum of 365° K at 60 km., then falls rapidly to almost 180° K at 80 km. after which it again rises. The work of Pekeris (1937) on atmospheric oscillations is in general agreement with this description.

Some further confirmation is given by the measurements of the scale heights of the *D* and *E* layers (§§ 3·1, 3·2) which show the temperatures to be about 200° K at 70 km. and 400° K at 120 km. It must be remembered, however, that the interpretation of the experimental results is not at all simple, thus:

(i) The scale height derived from them refers to the atmospheric constituent that gives rise to the ionization. For the *E* layer, account must be taken of the important influence of dissociation on the distribution of atomic and molecular oxygen.

(ii) The values obtained depend on rather uncertain assumptions. In the D layer the determination is based on the magnitude of the diurnal variation of the altitude. This may be complicated by thermal and other effects—it will be noted in this connexion that the phase of the corresponding seasonal variation is peculiar (Budden, Ratcliffe & Wilkes 1939). The scale height in the E layer is derived from the structure of the ionization but no allowance is made for the thickening (as compared with a simple Chapman layer (§ 4·4)) that occurs if the absorption cross-section involved is not independent of the frequency of the ionizing radiation.

Little information is available on the temperatures at great altitudes, but the probable high density indicates that they must be considerable—at least 500° K. The scale heights found for the F layers (§ 3·3) suggest that in the region 200–300 km. the temperatures are from 1000 to 2000° K. However, it must be emphasized that the huge ionization thickness upon which these figures are based may well be at least partially accounted for by the composite nature of the individual layers or by the effective recombination coefficient being pressure dependent.

Spectroscopic investigations may eventually provide further data on this subject. The measurements available at present do not appear to favour very high temperatures; they cannot, however, be regarded as conclusive.

2·3. *Density*

Meteors give the most important direct information on the density of the upper atmosphere. Taking the mean molecular weight as 28·8, Whipple (1943) finds that at 60 km. the number of particles/c.c. is 8×10^{15}, at 80 km. $1{\cdot}4 \times 10^{15}$, at 100 km. 9×10^{13}, and at 120 km. 6×10^{12}. Integration using the estimated temperature distribution leads to slightly lower densities, but the discrepancy is not great if allowance is made for an increase in the scale height due to dissociation of O_2. The observed collision frequency in the E region is not inconsistent with these values.

It is not yet possible to obtain any reliable indications above the E layer. The evidence from the measurements of the scale heights suggests very tentatively 1×10^{11} particles/c.c. at 220 km. (the F_1 layer) and 2×10^{10} particles/c.c. at 300 km. (the F_2 layer). Though these estimates are perhaps rather too large, difficulties would be introduced by accepting very much lower values—for example, abnormally high absorption cross-sections would be necessitated to explain the production of the upper ionized layers (§ 4·4). There are several determinations of the collision frequency in the F regions (cf. § 3·3), but, as it is probable that electron-electron collisions provide the major contribution at these levels, no information concerning the gas pressure there can be derived from them.

It is clear that the atmosphere is still far from fully studied. The lack of reliable data makes difficult the application of the desirable quantitative tests on proposed theories and so leaves much scope for speculation. It is imperative that the present methods of exploration be extended as far as possible and that the newer techniques (involving for example modern search-light beams and rocket projectiles) be fully exploited.

3. PROPERTIES OF THE IONIZED LAYERS

The observed properties of the *D*, *E* (normal and sporadic), F_1 and F_2 (*F*) layers* are now summarized for later reference—attention being confined mainly to the results obtained by a few groups of workers and no attempt being made to give a complete account of the vast amount of data now available.

3·1. *The D layer*

The most recent study of the *D* layer has been made by Budden *et al.* (1939) using the Rugby transmitter. From the reflexion of long radio waves they established the existence, under normal conditions, of a weakly ionized layer at about 70 km. The diurnal variations were consistent with a scale height of about 6 km. The seasonal behaviour was, however, anomalous in that the layer appeared lower in winter than in summer (§ 4·4).

Radio fade-outs are caused by a great increase of the ionization which produces strong absorption owing to the high electron collision frequency. Dellinger (1937) has shown that there is a correlation between their occurrence and the existence of solar chromospheric eruptions. It is not yet certain whether the normal and the highly ionized layers are coincident, and whether they are distinct from or merely downward promulgations of the *E* layer. Much experimental work has still to be done.

3·2. *The E layers (normal and sporadic)*

Measurements on the noon height of the *E* layer show it to be just over 100 km. throughout the year at Huancayo 1938 (Wells & Stanton 1938*a*, *b*, 1939), 110 km. in summer and 125 km. in winter at Washington 1935 (Gilliland, Kirby, Smith & Reymer 1936), and 120 km. in summer and 134 km. in winter at Slough 1936 (Appleton & Naismith 1940). The corresponding values of the maximum electron density n_e† are respectively, $2{\cdot}2\times10^5$/c.c., $1{\cdot}8\times10^5$/c.c. and $1{\cdot}2\times10^5$/c.c., and $1{\cdot}7\times10^5$/c.c. and $1{\cdot}0\times10^5$/c.c. The diurnal variation of n_e follows the quadratic law

$$dn_e/dt = q - \alpha n_e^2, \tag{3}$$

q being the effective rate of electron production and α the effective recombination coefficient. Throughout a large part of the day dn_e/dt can be neglected so that

$$n_e = (q/\alpha)^{\frac{1}{2}} \tag{4}$$

$$= (q_0/\alpha)^{\frac{1}{2}}(\cos\chi)^{\frac{1}{2}}, \tag{5}$$

---

* Other layers have been suggested below the *D* layer (the *C* layers) (Mitra & Bhar 1936; Colwell & Friend 1936; Watson-Watt, Bainbridge-Bell, Wilkins & Bowen 1936) and above the F_2 layer (the *G* layer) (Kirby, Berkner & Stuart 1934; Maeda 1937), but it is doubtful if the reflexions suggesting the *C* layers are produced by ionization (Berkner 1939), and those suggesting the *G* layer are probably produced by multiple reflexions by the ordinary layers (Eckersley 1939).

† The electron densities given in this paper are derived from the critical frequencies without the Lorentz polarization correction.

where χ is the solar zenith angle. Comparison of observations at different latitudes and times of year yields the important result that α is a constant independent of the gas density in the range of E layer height (Appleton, Naismith & Ingram 1937). Several estimates have been made of the absolute value of α: Appleton (1937) gives 10^{-8} c.c./sec., Wilkes* (1939) $1{\cdot}2 \times 10^{-8}$ c.c./sec., Hulburt (1939) 2×10^{-8} c.c./sec., Bhar (1939) 4×10^{-9} c.c./sec., and Higgs† (1942) $1{\cdot}0 \times 10^{-8}$ c.c./sec. At night n_e falls to a low value of order 10^4/c.c. There appears to be some irregularity in the rate of decay. Various values of α (nocturnal) have been given: 2×10^{-9} c.c./sec. (Appleton 1937), 4×10^{-9} c.c./sec. (Wilkes 1939), and 1×10^{-9} c.c./sec. (Bhar 1939). These are probably too small because of the neglect of the effects of sources of ionization, and it seems best to assume that the true night value is the same as the day value.

Very little work has been done on the determination of the scale height of the layer. Appleton (1939 *a*), from the variation of the equivalent height with the frequency of the probing radio wave gives it to be 11·4 km. for a summer noon. Application of the formula‡

$$h_m = H \log (n_0 A H / \cos \chi) \tag{6}$$

to the small diurnal variations of h_m give H to be markedly less than 10 km., but application to the considerable annual and latitudinal variations suggests (though the results are not very consistent) a much larger value. It is clear that there are complicating factors for which proper allowance is not made. Attempts at the determination of the electron collision frequency ν have also been somewhat neglected since the early calculations of Bailey & Martyn (1935) on the interaction of radio waves led to the value 2×10^5/sec. Direct radio evidence on the negative-ion electron ratio λ under various conditions is still incomplete. Systematic measurements in these three directions would be very useful.

Appleton & Naismith (1939) have examined the variation of the ionization at Washington, Slough and Tromso during the years 1931–8. They find that n_e increases by a factor of 1·50 from sunspot minimum to sunspot maximum.

The sporadic E layer takes the form essentially of patches of ionization which may be of considerable density—often apparently greater than 10^6 electrons/c.c. It is located in the region of the normal layer. Appleton & Naismith (1940) (working at Slough 1934–8) determined the height as 113 km. in summer and 130 km. in winter. Most of the experimental work so far has been directed towards investigating the general properties of the layer so as to get some indication as to its origin. The ionization is found to be more pronounced during the day than during the night, and during the summer than during the winter. It appears to increase slightly from sunspot minimum to sunspot maximum, but this may not be significant. The effects at high latitudes are the most marked, and there is a definite correlation with magnetic activity (Appleton *et al.* 1937; Berkner 1939; Appleton & Naismith 1940).

---

* His analysis is based on the work of Kirby, Gilliland & Judson (1936) (solar eclipse measurements) and of Best, Farmer & Ratcliffe (1938).

† Solar eclipse measurements.

‡ See § 4·4.

Pierce (1941) has carried out some useful preliminary work on the rate of recombination in individual clouds. Further exploratory research has still to be done, but eventually valuable information should be obtained. The effects of altitude are of special interest.

3·3. *The F_1 and F_2 (F) layers*

The F_1 and F_2 layers have separate identities only when the solar zenith angle is high. At other times they merge together to form the F layer. Typical measurements of the height and maximum electron density are given in table 1.

TABLE 1. TYPICAL MEASUREMENTS OF THE HEIGHT AND MAXIMUM ELECTRON DENSITIES IN THE F_1 AND F_2 (F) LAYERS

	height (km.)	max. electron density (cm.$^{-3}$)
F_1, Huancayo, midday, June 1938 (Wells & Stanton 1938*a*, *b*)	208	$3{\cdot}7 \times 10^5$
F_1, Huancayo, midday, Dec. 1938 (Wells & Stanton 1939)	204	$3{\cdot}7 \times 10^5$
F_2, Huancayo, midday, June 1938 (Wells & Stanton 1938*a*, *b*)	325	$1{\cdot}1 \times 10^6$
F_2, Huancayo, midday, Dec. 1938 (Wells & Stanton 1939)	298	$1{\cdot}7 \times 10^6$
F, Huancayo, midnight, June 1938 (Wells & Stanton 1939)	223	$6{\cdot}5 \times 10^5$
F, Huancayo, midnight, Dec. 1938 (Wells & Stanton 1939)	326	$7{\cdot}4 \times 10^5$
F_1, Washington, midday, June 1935 (Gilliland *et al.* 1936)	236	$2{\cdot}5 \times 10^5$
F_1, Washington, midday, Dec. 1935 (Gilliland *et al.* 1936)	219	$2{\cdot}1 \times 10^5$
F_2, Washington, midday, June 1935 (Gilliland *et al.* 1936)	382	$4{\cdot}9 \times 10^5$
F_2, Washington, midday, Dec. 1935 (Gilliland *et al.* 1936)	243	$1{\cdot}4 \times 10^6$
F, Washington, midnight, June 1935 (Gilliland *et al.* 1936)	296	$2{\cdot}7 \times 10^5$
F, Washington, midnight, Dec. 1935 (Gilliland *et al.* 1936)	303	$2{\cdot}5 \times 10^5$

Observation shows the behaviour or the F_1 layer to be quite normal—in particular the quadratic law (3) is obeyed and the recombination coefficient is constant (Appleton & Naismith 1935). The F_2 layer is much less regular. It exhibits some very remarkable characteristics. These have been discussed in detail by Appleton & Naismith (1935), Goodall (1939) and Berkner (1939). Certain of the more interesting features, however, may be mentioned here. The diurnal variation of the maximum electron density is peculiar in that there is a pronounced noon depression, during the summer at northern latitudes and throughout the year at equatorial latitudes. The annual variation can be analysed into seasonal and non-seasonal components of approximately equal amplitudes, the former with a maximum in local winter and the latter with a maximum round December.

Higgs* (1942) estimated the recombination coefficient α to be $8{\cdot}5 \times 10^{-9}$ c.c./sec. in the F_1 layer. A rather smaller value (possibly 4×10^{-9} c.c./sec.) seems, however, to be more probable. Further measurements are needed. Several determinations have been made for the F_2 layer. Appleton (1937) gives α as $8{\cdot}7 \times 10^{-11}$ c.c./sec. for a summer day, $8{\cdot}1\times10^{-11}$ c.c./sec. for an equinox day and a higher (but unmeasured) value for a winter day. Bhar (1939) gives 4×10^{-11} c.c./sec. for a summer day, Higgs* (1942) 4×10^{-11} c.c./sec. for an equinox day, and Mohler (1940) 9×10^{-11} c.c./sec. as an average over 20 months 1937–9 (sunrise and sunset). During the night α appears to increase to about 3×10^{-10} c.c./sec. (Appleton 1937). Mohler (1940), however, finds values ranging from about 5×10^{-11} to 5×10^{-10} c.c./sec. and points out that there is evidence suggesting that an attachment rather than a recombination law is being obeyed.

---

* Solar eclipse measurements.

Only a few isolated measurements have been made of the scale heights. Appleton's method leads to values of 70 km. for the F_2 layer in summer and to 40 km. for the F layer in winter. A value of about 30 km. seems plausible for the F_1 layer. The effects of temperature changes, etc., are too important to enable the formula (8) to be used successfully. A considerable number of investigations have been carried out relating to the electron collision frequency. The following are among the results that have been obtained:

F_1, October day 1934 at Chelmsford, $\nu = 3{\cdot}6\times10^3$/sec. (Eckersley 1935)
F_2, November day 1934 at Cambridge, $\nu = 1{\cdot}2\times10^3$/sec. (Farmer & Ratcliffe 1935)
F, November night 1934 at Cambridge, $\nu = 2{\cdot}1\times10^3$/sec. (Farmer & Ratcliffe 1935)
F, March night 1936 at Allahabad, $\nu = 1{\cdot}2\times10^4$/sec. (Toshniwal, Pant & Bajpai 1936)

There is no full understanding of the factors leading to these various values, and further systematic work, such as that of White & Brown (1935), is required before the matter can be elucidated properly. It is probable that the ionization in the layers is predominantly electronic.

During the sunspot cycle the maximum electron density in the F_1 layer varies by a factor 1·56 (Appleton & Naismith 1939), while that in the F_2 layer by a factor of about 4 (Appleton 1939*b*).

3·4. *Model ionosphere*

The fundamental theory of the ionosphere is not yet at a stage at which it can explain the full observational complexity, and for purposes of initial discussion we take a simplified representative model as follows:

E layer

Altitude = 120 km.
Maximum electron density = $1{\cdot}5\times10^5$/c.c. (day), 1×10^4/c.c. (night).
Recombination coefficient = 1×10^{-8} c.c./sec. (day and night).
Scale height = 10 km.
Gas density = 6×10^{12}/c.c.
Temperature = 500° K.

---

* Solar eclipse measurements.

F_1 layer

Altitude = 220 km.
Maximum electron density = $2 \cdot 5 \times 10^5$/c.c.
Recombination coefficient = 4×10^{-9} c.c./sec.
Scale height = 30 km.
Gas density = 1×10^{11}/c.c.
Temperature = 1000° K.
} merged with F_2 layer at night.

F_2 layer

Altitude = 300 km.
Maximum electron density = $1 \cdot 0 \times 10^6$/c.c. (day), $2 \cdot 5 \times 10^5$/c.c. (night).
Recombination coefficient = 8×10^{-11} c.c./sec. (day), 3×10^{-10} c.c./sec. (night).
Scale height = 70 km.
Gas density = 2×10^{10}/c.c.
Temperature = 2000° K.

The uncertainties that exist (particularly in connexion with the gas densities) must be borne in mind.

4. PRODUCTION OF ATMOSPHERIC IONIZATION

4·1. *Introduction*

It is clear, from the properties of the layers summarized in the preceding section, that solar radiation is responsible for all the upper atmospheric ionization except perhaps for part of the F_2 layer and for the sporadic E layer. The peculiar unexplained variations which the F_2 layer exhibits throws doubt on the complete control of its behaviour by the sun,* but there is sufficient evidence to prove that, even here, a large proportion of the ionization arises from this source. The sporadic E layer may be due to a variety of causes, but apart from the ionization that is undoubtedly produced by meteors (which is best regarded as distinct from the true sporadic E layer) a solar origin is the most probable here also.

As the change of ionization during eclipses gives no indication of appreciable effects from corpuscular radiation in the E and F_1 layers (Appleton & Chapman 1935), it seems preferable to assume that these are produced by electro-magnetic radiation. The observational data of course do not exclude the possibility of ionization by very fast corpuscles. It is difficult to see how any direct evidence regarding these can be obtained. Though the position is not so clear for the F_2 layer it is again taken that electro-magnetic radiation is the main cause of ionization (Pierce, Higgs & Halliday 1940; Higgs 1942). In view of the complex diurnal variation of the layer the ordinary simple eclipse effect can perhaps scarcely be expected. It may be significant that in general the most definite evidence (which shows that the ionizing radiation has a very high velocity) has been obtained at low solar altitude. The correlation between solar chromospheric eruptions and radio fade-outs favours electro-magnetic radiation as the origin also of the *D* layer, but once again very fast corpuscles are not excluded. In this connexion the suggestion of Chapman that a burst of corpuscular radiation might be detected by its absorption spectrum is of interest. There is much more doubt concerning the sporadic *E* layer. Neglecting meteoric ionization, the behaviour suggests charged solar corpuscles as the principal source (Appleton & Naismith 1940), but no completely convincing theory has yet been put forward.

---

* Eckersley (1940) suggested that the non-seasonal daytime variation of the F_2 layer might be due to some form of radiation from outer space (possibly correlated with the Jansky noise). Measurements of the night-time variations were available for the Northern hemisphere alone, and these were in agreement with his hypothesis in that the ionization was a maximum in summer (corresponding to the daytime maximum in winter). However, examination of recent measurements for the Southern hemisphere shows that the night-time variation is probably only seasonal.

4·2. *Solar electro-magnetic radiation*

It is usually assumed that the normal electro-magnetic radiation from the sun is that characteristic of a black body at temperature 6000° K. Little else can be done in view of the present state of knowledge of solar physics, but it must be realized that the error involved may be considerable. A summary of the meagre evidence available has been given by Hunter (1943). This suggests that the radiation beyond the ultra-violet region is in excess and may contain very strong emission lines (in particular $\lambda 1215$ A of H and $\lambda 584$ A of He). These features are probably especially pronounced during the transitory chromospheric eruptions associated with radio fade-outs.

4·3. *The ionization potentials and continuous absorption cross-sections of the main atmospheric gases*

Table 2 gives the energies required for the ionization of O, O_2 and N_2 together with the flux of solar quanta available.

TABLE 2

gas	energy required for ionization (eV)	flux of solar quanta available* (/cm.^{2}/sec.)
O_2	12·2 (first level)	$1{\cdot}0 \times 10^{10}$
O	13·5 (first level)	$9{\cdot}3 \times 10^{8}$
N_2	15·5 (first level)	$2{\cdot}7 \times 10^{7}$
O_2	16·1 (second level)	$8{\cdot}1 \times 10^{6}$
O_2	16·9 (third level)	$2{\cdot}1 \times 10^{6}$
O	16·9 (second level)	$2{\cdot}1 \times 10^{6}$
O_2	18·2 (fourth level)	$1{\cdot}8 \times 10^{5}$
O	18·5 (third level)	$1{\cdot}0 \times 10^{5}$
N_2	18·7 (second level)	$7{\cdot}4 \times 10^{4}$

The continuous absorption cross-section of atomic oxygen has been calculated using quantum theory by several groups of workers (Saha & Rai 1938; Bates *et al.* 1939; Yamanouchi & Katani 1940). The results are shown in table 3. They should be quite accurate, as interference effects in the evaluation of the matrix element are unlikely to be serious (Bates & Massey 1941).

---

* The sun is taken to be equivalent to a black body at 6000° K. (§ 4·2).

TABLE 3. CONTINUOUS ABSORPTION CROSS-SECTION FOR ATOMIC OXYGEN

energy range of incident quanta (eV)	continuous absorption cross-section (cm.^{2})
13·5–16·9	$4{\cdot}5 \times 10^{-18}$
16·9–18·5	11×10^{-18}
18·5–*c.* 25	16×10^{-18}

A certain amount of investigation has been carried out relating to the photo-ionization of molecular oxygen and nitrogen.* From the path lengths for absorption it is possible to estimate the cross-section involved. Approximate values are: for O_2, 10^{-20} cm.^{2} at the first ionization potential and 10^{-16} to 10^{-17} cm.^{2} at the higher ionization potentials; and for N_2, 10^{-17} cm.^{2} at the first ionization potential and 10^{-16} cm.^{2} at the second ionization potential. Further measurements are very desirable.

Finally, it may be mentioned that the continuous absorption cross-sections based on Kramer's formula (that are sometimes quoted) are unreliable (Bates 1939).

4·4. *Ionized layer formation*

In the E and probably the F_1 layers the recombination coefficient is independent of the gas pressure so that the maximum ionization density occurs at the level of maximum ion production. In the D and F_2 layers the situation may be more complex.

For a homogeneous atmosphere Chapman (1931) has shown that $I(h, \chi)$, the rate of production of ions at altitude h by solar radiation of zenith angle χ, is given by

$$I(h,\chi) = S_0 nA \exp(-nAH \sec\chi), \tag{7}$$

$$n = n_0 \exp(-h/H), \tag{8}$$

where H is the scale height and n the density of the atmosphere, S_0 the number of quanta incident/cm.^{2}/sec. and A the absorption cross-section. I has a maximum,

$$I_m = S_0 \cos\chi/(H \exp 1), \tag{9}$$

where the altitude is $$h_m = H \log(n_0 AH/\cos\chi), \tag{10}$$

and the density is $$n_m = \cos\chi/AH. \tag{11}$$

The total rate of production of ions is given, per cm.^{2} column, by

$$Q = \int I dh$$
$$= S_0 \cos^2\chi. \tag{12}$$

Similar expressions can be obtained for a non-homogeneous atmosphere containing several constituents but they are somewhat cumbersome, and for immediate purposes it is sufficient to note that the level of maximum ion production is raised if the radiation is absorbed by some process other than that leading to ionization.

---

* The work is summarized by Chapman & Price (1936) and Price (1943).

4·5. *The E, F_1 and F_2 layers*

It can readily be seen that it is improbable that quanta with enough energy to ionize atomic oxygen can penetrate sufficiently deep into the atmosphere to be responsible for the *E* layer. Taking the continuous absorption cross-section for quanta with energy between 13·5 and 15·5 eV as $4{\cdot}5 \times 10^{-18}$ cm.^{2} and the scale height as 30 km. (corresponding to the F_1 layer), the maximum absorption occurs where the density of oxygen atoms is $7 \times 10^{10} \cos\chi$/c.c. This is well above the *E* layer. Since the intensity of the radiation diminishes very rapidly below the level of maximum absorption (at distances H, $2H$ and $3H$ below this level it is reduced by factors of $2{\cdot}5 \times 10^{-2}$, 2×10^{-4} and 10^{-9} respectively from the incident value) it is very weak indeed by the time the *E* layer is reached. Quanta of energy greater than 15·5 eV are even less likely to reach the *E* layer because of their stronger absorption (by both O and N_2).

The formation of the *E* layer must, on this basis, be ascribed to ionization produced by radiation with quantum energy less than 13·5 eV. Reference to table 2 shows that the only possibility is the ionization of O_2 at 12·2 eV. Accepting for the moment that this is correct it is natural to associate the F_1 layer with absorption by O at its first ionization potential (13·5 eV), and the F_2 layer with absorption by O at its higher ionization potentials (16·9 and 18·5 eV) and with absorption by N_2 at its ionization potentials (15·5 and 18·7 eV). Though this interpretation has certain weaknesses it is attractive.

According to experiment the continuous absorption cross-section for O_2 at its first ionization potential (12·2 eV) is of order 10^{-20} cm.^{2}, so that as the scale height of the *E* layer is 10 km. the density of O_2 at the level of maximum absorption should be about $10^{14} \cos\chi$/c.c. This is greater than that indicated by table 1, but the discrepancy may partially arise from non-ionizing absorption.* A further difficulty is that the altitude of the transition region between molecular and atomic oxygen would have to be greater than is generally accepted (particularly as the *E* layer if produced as suggested would tend to be *below* rather than *in* this region).

The view has often been put forward that the relatively small thickness of the layer is evidence in support of the proposed origin. This argument is based on a misconception—the measured scale height is actually rather larger than might be expected for O_2 on the Wulf-Deming theory (§ 2·1) and can be associated quite naturally with the main atmosphere.

At the level of maximum absorption in the F_1 layer the density of O should be $7 \times 10^{10} \cos\chi$/c.c. (the continuous absorption cross-section being $4{\cdot}5 \times 10^{-18}$ cm.^{2} and the scale height being 30 km.). This is in satisfactory agreement with expectation.

Unless the energy distribution of the incident radiation is very different from that assumed (§ 4·2), the most important of the contributions to the F_2 layer is that from N_2 at its first ionization potential (15·5 eV). The continuous absorption cross-section of 10^{-17} cm.^{2} and scale height of 70 km. gives the density of N_2 at the level of maximum absorption as $1\cdot 4 \times 10^{10} \cos\chi$/c.c.—again of the correct order.

---

* See, however, § 6·1 (Equilibrium of *E* layer).

Table 4 compares the flux of solar quanta available for the various layers with the observed rate of production of ionization, per cm.^{2} column. It is interesting to note that the suggestion of non-ionizing absorption is again apparent in connexion with the E layer. The data for the F_1 and more particularly the F_2 layer indicates perhaps that the 6000° K black-body model under-estimates the intensity in the high-energy region of the spectrum.

TABLE 4

layer	flux of solar quanta available (cm.^{2}/sec.)	observed rate of production of ionization (/cm.^{2} column/sec.)
E (O_2, first ionization potential)	$1\cdot 0 \times 10^{10}$	6×10^8
F_1 (O, first ionization potential)	$9\cdot 3 \times 10^8$	2×10^9
F_2 (O, second and third ionization potentials; N_2, first and second ionization potentials)	$2\cdot 7 \times 10^7$	$1\cdot 5 \times 10^9$

Other theories on the production of the layers have been put forward by Bhar (1938), Mitra, Bhar & Ghosh (1938) and by Wulf & Deming (1938). The former ascribe the E layer to ionization of O_2 at its second ionization potential (16·1 eV), the F_1 layer to the ionization of N_2 at its second ionization potential (18·7 eV), and the F_2 layer to ionization of O at its first ionization potential (13·5 eV). The latter also ascribe the E layer to ionization of O_2 but at its first ionization potential (12·2 eV), and the F_1 and F_2 layers to ionization of N_2 at its second (18·7 eV) and first (15·5 eV) ionization potentials respectively. Both theories therefore fail to take into account the effect of the strong absorption by O of radiation between 13·5 and (at least) 25 eV. Thus, in the theory of Bhar and his co-workers, if the F_2 layer is formed by O at its first ionization potential (13·5 eV), then there will certainly not be sufficient radiation of 16·1 or 18·7 eV quantum energy penetrating that region to produce the lower two layers. Wulf & Deming essentially ignore O altogether. It may also be remarked that for both theories the magnitude of the observed changes in the different layers with sunspot cycle would require the variation of the intensity of the solar radiation to have a peculiar frequency dependence—the sensitivity of the layers to the sunspot cycle increases in the order $E-F_1-F_2$. Although the theory initially discussed is not without defects it does not appear to encounter such serious difficulties as these alternatives.

The only remaining purely electro-magnetic radiation theory that seems even at all feasible is that the E layer should be attributed to O at its first ionization potential (13·5 eV) and the F_1 and F_2 layers to a combination of N_2 at its first and second ionization potentials (15·5 and 18·7 eV respectively) and of O at its second and third ionization potentials (16·9 and 18·5 eV respectively). As already pointed out it is not easy to reconcile such an origin for the E layer with the quantal calculations on atomic oxygen. Mohler (1940) has proposed that the F layers arise from a single process—one being formed at the level of maximum ion production in the normal manner and the other being formed at a much greater height owing to the effective recombination coefficient being pressure dependent. The observed difference in the coefficients is certainly sufficient to give rise to two layers in this way. While the quantitative evidence is inconclusive it is satisfactory to note that the ratio of the rate of production of ionization/unit gas density in F_2 to that in F_1 is 1·6, which is not significantly lower than the predicted value e (2·7).

Finally, it may be mentioned that Gauzit (1943) is incorrect in suggesting that considerable ionization at the level of the E layer might result from the action

$$O(^1S) + O(^1S) \rightarrow O_2^+ + e. \tag{13}$$

Numerical examination shows at once that the yield is negligible owing partly to the number of metastable atoms that exist even during the day being small (§ 7·1·1) and partly to the collision probability involved being low (certainly much less than unity). The possibilities of a theory involving corpuscular radiation may have to be explored later.

Spectroscopic evidence on the products of ionization would be very useful in spite of the complications by the effects of electron transfer reactions (§ 5·4). The prospect of observing absorption in the solar spectrum is unfortunately slight, but the twilight emission (§ 7·2·2), though probably negligible for O^+ (on account of the rarity of quanta with enough energy (14·8 eV) to raise the ion through its first allowed transition), may be appreciable for O_2^+ (which requires 4·7 eV for the Second Negative system, $A\,^2\Pi \rightarrow X\,^2\Pi$) and for N_2^+ (which requires 3·1 eV for the First Negative system, $B\,^2\Sigma \rightarrow X\,^2\Sigma$).* N_2^+ bands have actually been reported (§ 7·1·3) so that hopes are raised that eventually information will be obtained concerning its height distribution both at sunrise and sunset and that it will prove feasible to examine the intensity variation (particularly throughout the sunspot cycle) for correlations with the properties of the layers.

4·6. *The D and sporadic E layers*

The quantitative properties of the D and sporadic E layers have not been determined sufficiently well to justify an attempt at a detailed analysis. However, as the former is probably produced by electro-magnetic radiation any theory of the upper layers should leave some provision for it. Among the obvious sources is the ionization either of one of the rarer atmospheric constituents or (if this does not occur in the normal E layer (§ 4·4)) of O_2 at 12·2 eV. There is insufficient evidence to decide between the various possibilities.

---

* Calculations on the intensity to be expected can be made, using formula (88), but it is only possible to obtain the approximate order of magnitude owing to the lack of reliable data —the values of the radiative lifetimes of the excited ions, for example, are uncertain. However, it can at least be stated that there is no evidence of discrepancy with observation.

5. RECOMBINATION, ATTACHMENT, DETACHMENT AND TRANSFER REACTIONS

In the previous section the production of the ionized layers has been discussed. To proceed further and consider their behaviour it is necessary to know the rates of the various collision processes of importance. These can be classified into four categories—recombination of oppositely charged particles, attachment of electrons to neutral particles, detachment of electrons from negative ions, and electron transfer between neutral particles and positive or negative ions.

5·1. *Recombination*

Recombination can be divided into two main types, electronic recombination and ionic recombination.

5·1·1. *Electronic recombination*

The main reactions are:

$$\left.\begin{array}{lcr} & O^+ + e \rightarrow O + h\nu; & (14a) \\ \text{similarly for} & O_2^+ \ (14b) \quad \text{and} \quad N_2^+. & (14c) \end{array}\right\} \begin{array}{l}\text{radiative}\\ \text{recombination}\end{array}$$

$$\left.\begin{array}{lcr} & \left.\begin{array}{c} O^+ + e \rightarrow O'^{(D)} \ \text{(doubly excited)}, \\ O'^{(D)} \rightarrow O + h\nu; \end{array}\right\} & (15a) \\ \text{similarly for} & O_2^+ \ (15b) \quad \text{and} \quad N_2^+ & (15c), \end{array}\right\} \begin{array}{l}\text{di-electronic}\\ \text{recombination}\end{array}$$

$$\left.\begin{array}{lcr} & O_2^+ + e \rightarrow O' + O''; & (16a) \\ \text{similarly for} & N_2^+. & (16b) \end{array}\right\} \begin{array}{l}\text{dissociative}\\ \text{recombination}\end{array}$$

The rate of the reaction (14*a*) has been calculated (Bates *et al.* 1939; Massey & Bates 1943) when the resulting atom is in the ground and any excited state. The total recombination coefficient is found to be $3{\cdot}7 \times 10^{-12}$ c.c./sec. at 250° K, and $1{\cdot}5 \times 10^{-12}$ c.c./sec. at 1000° K. No detailed calculations have been made for (14*b*) and (14*c*), but they will give rise to much the same total recombination coefficient as (14*a*).

The di-electronic process (15*a*) has been discussed by Massey & Bates (1943). It appears that the recombination coefficient involved is not greater than 10^{-12} c.c./sec.; that for (15*b*) and (15*c*) is also likely to be small.

The importance of the dissociative reactions (16*a*) and (16*b*) is rather difficult to assess. Similar reactions between electrons and neutral molecules leading to attachment proceed comparatively rapidly.* Detailed investigations are clearly required.

It is appropriate to mention here an unexplained set of results relating to electronic recombination. Experiments have been carried out to measure the coefficient for recombination between electrons and positive atomic ions in argon (Kenty 1928), mercury (Mohler 1937) and caesium (Mohler 1937). For all three elements the recombination coefficient was determined as of the order 10^{-10} c.c./sec. This is about 100 times larger than would have been expected from radiative recombination and (in spite of the many contributory doubly excited states possible in heavy atoms such as those investigated) is probably too great to be accounted for by di-electronic recombination.* It is important to discover whether it is a coincidence that the recombination coefficient observed in the F layers is also about 10^{-10} c.c./sec. or whether some process has been overlooked.

---

* The dissociative attachment coefficient to O_2 is 10^{-10} c.c./sec. (Massey 1937).

5·1·2. *Ionic recombination*

There is ample evidence (Massey 1938) of the existence of O^- and O_2^-,† but N_2^- does not appear to be stable. The ionic recombination reactions of importance may thus be written

$$\left.\begin{array}{lcl} & O^- + O^+ + M \rightarrow O + O + M; & (17a)\\ \text{similarly for} & O_2^+\ (17b) \quad \text{and} \quad N_2^+ & (17c)\\ & O_2^- + O^+ + M \rightarrow O_2 + O + M; & (18a)\\ \text{similarly for} & O_2^+\ (18b) \quad \text{and} \quad N_2^+ & (18c) \end{array}\right\}\begin{array}{l}\text{three-body}\\ \text{recombination}\end{array}$$

$$\left.\begin{array}{lcl} & O^- + O^+ \rightarrow O' + O''; & (19a)\\ \text{similarly for} & O_2^+\ (19b) \quad \text{and} \quad N_2^+ & (19c)\\ & O_2^- + O^+ \rightarrow O_2' + O''; & (20a)\\ \text{similarly for} & O_2^+\ (20b) \quad \text{and} \quad N_2^+ & (20c) \end{array}\right\}\begin{array}{l}\text{mutual}\\ \text{neutralization}\end{array}$$

The three-body reactions (17), (18) have been investigated by Thomson (1924), and the correctness of his theory has been established by experiments carried out by several workers (Gardner 1938; Sayers 1938). Under ionospheric conditions the recombination coefficient is proportional to the pressure and is inversely proportional to the three-halves power of the temperature. Neglecting the relatively unimportant correction which allows for the particular ions concerned it is given simply by $4 \times 10^{-5} p/T^{\frac{3}{2}}$ c.c./sec. (p the pressure in mm. Hg, and T the temperature in ° K).

An estimate of the possible rates of mutual neutralization ((19), (20)) has been made by Bates & Massey (1943).‡ Although it was not found possible to derive a theoretical value for the recombination coefficient in any particular case, it was shown that under certain conditions depending on the detailed properties of the atoms or molecules concerned a large value, of order 10^{-7}–10^{-8} c.c./sec. but not much greater, could be attained. For most reactions it is probable that the conditions are satisfied for some pair of excited products. An experimental investigation is desirable.

Radiative recombination of ions is never important in the upper atmosphere, the recombination coefficient being extremely small.

---

* Experiments with light atoms such as helium, where di-electronic recombination is certainly small, would be useful.

† Lozier (1934) measured the detachment energy of O^- as 2·2 eV, and this value is in general accepted in the present paper. Recent measurements of Vier & Mayer (1944), however, yield a value of 3·1 eV. The detachment energy of O_2^- is rather uncertain. Loeb (1935) gives 0·07–0·34 eV and Bloch & Bradbury (1935) 0·07–0·17 eV, but Bates & Massey (1943) favour a much higher value of about 1 eV.

‡ They only discussed the reaction (19a), but their conclusions can be applied generally.

5·2. *Attachment*

The following are the principal attachment reactions to be considered

$$\left.\begin{array}{lcr} & \mathrm{O}+e\rightarrow\mathrm{O}^-+h\nu, & (21a) \\ \text{similarly for} & \mathrm{O}_2, & (21b)\end{array}\right\}\begin{array}{l}\text{radiative}\\ \text{attachment}\end{array}$$

$$\left.\begin{array}{lcr} & \mathrm{O}+e\rightarrow\mathrm{O}^{-\prime(D)}\ \text{(doubly excited)}, & (22a) \\ & \mathrm{O}^{-\prime(D)}\rightarrow\mathrm{O}^-+h\nu; & \\ \text{similarly for} & \mathrm{O}_2 & (22b)\end{array}\right\}\begin{array}{l}\text{di-electronic}\\ \text{attachment}\end{array}$$

$$\mathrm{O}_2+e\rightarrow\mathrm{O}+\mathrm{O}^-. \qquad (23b)\ \text{dissociative attachment}$$

Radiative attachment to atomic oxygen (21*a*) has been investigated in detail by Bates & Massey (1943). They find that the attachment coefficient depends markedly on the binding energy of the ^{2}P term of the $(1s)^2\,(2s)^2\,(2p)^4\,3s$ configuration of O$^-$. If this is zero, resonance effects occur and the attachment coefficient is very large—6×10^{-14} c.c./sec. at 250° K and $1{\cdot}5\times10^{-14}$ c.c./sec. at 1000° K. It is more probable, however, that the excited state is unstable and that the attachment coefficient is only about $1{\cdot}1\times10^{-15}$ c.c./sec. and practically independent of the temperature. In these calculations the electron affinity of O ions was taken as 2·2 eV as derived from Lozier's collision experiments (1934). If the recent value of 3·1 eV as measured thermochemically by Vier & Mayer (1944) is taken instead, the attachment coefficients quoted above would all have to be multiplied by 2·8.

The reaction (21*b*) has similar possibilities to (21*a*) but no complete theory has been given.

The di-electronic process (22*a*) has been studied by Bates & Massey (1943). It is unlikely to be appreciable compared with (21*a*) unless O$^-$, $(1s)^2\,(2s)\,(2p)^6\,{}^2$S lies within about 0·25 eV of O $(1s)^2\,(2s)^2\,(2p)^4\,{}^3$P. This does not appear at all probable. Analysis of the deep electronic states of O_2^- indicates that (22*b*) is also unimportant.

Dissociative attachment (23*b*) can only occur with electrons of energy greater than 3 eV (Lozier 1934; Hagstrum & Tate 1941), so that it need not be considered in ionospheric problems.

The attachment of very slow electrons to O$_2$ to form O_2^- at pressures of a few mm. Hg has been investigated by Bradbury (1933). The mechanism involved is not clear —the processes considered above seem quite inadequate. A discussion of the experimental results has been given by Bloch & Bradbury (1935) and by Bates & Massey (1943). It appears very probable that the major contribution to the observed attachment arises from a reaction which is pressure dependent, and that where p is the pressure (mm. Hg) the attachment coefficient is given by

$$(kp+\eta_0)\ \text{c.c./sec.} \quad (p<1), \qquad (24)$$

where k is probably at least $1{\cdot}5\times10^{-12}$ and η_0 is not greater than 10^{-14}.

5·3. *Detachment*

Here	$O^- + h\nu \rightarrow O + e;$	(25*a*)	photo detachment
similarly for	O_2^-.	(25*b*)	
	$O^- + h\nu \rightarrow O^{-\prime(D)}$ (doubly excited)	(26*a*)	di-electronic detachment
	$O^{-\prime(D)} \rightarrow O + e;$		
similarly for	O_2.	(26*b*)	
	$O^- + O \rightarrow O_2 + e;$	(27*a*)	associative detachment
similarly for	O_2 (27*b*) and N_2.	(27*c*)	
	$O_2^- + O \rightarrow O_3 + e;$	(28*a*)	
similarly for	O_2 (28*b*) and N_2.	(28*c*)	
	$O^- + X' \rightarrow O + X + e;$	(29*a*)	detachment by transfer of excitation
similarly for	O_2^-.	(29*b*)	

The rate at which the reaction (25*a*) occurs depends critically (as does that of the inverse, (21*a*)) on the stability of O^- $(1s)^2(2s)^2(2p)^4 3s\ ^2P$. For quanta with frequency distribution characteristic of a black body at 6000° K and dilution factor $5{\cdot}43 \times 10^{-6}$ the detachment rate is 0·85/negative ion/sec. if resonance conditions prevail, but the most likely value is only 0·35/negative ion/sec. (Bates & Massey 1943). These values are obtained on the assumption of 2·2 eV as the electron affinity of atomic oxygen. If the larger value of 3·1 eV is taken the detachment rates are reduced in the ratio 0·4 : 1.

No numerical values are available for the rate of (25*b*), but it would not be expected to differ much from (25*a*).

The reaction (26*a*) is the inverse of (22*a*). Its contribution to the detachment rate will only exceed that from (25*a*) in the unlikely circumstances of the energy of O^- $(1s)^2(2s)(2p)^6\ ^2S$ being in excess of O $(1s)^2(2s)^2(2p)^4\ ^3P$ by less than 0·25 eV (Bates & Massey 1943). (26*b*) can probably also be neglected.

Of the associative detachment processes, (27*b*), (27*c*), (28*a*) and (28*c*) are endothermic and so are unimportant in the upper atmosphere. (27*a*) and (28*a*), however, are exothermic and may play a vital role. Unfortunately, as yet there is no reliable means of studying them. Massey (1937) has estimated from the known rate of (23*b*) that the detachment coefficient for (27*a*) is 10^{-16}–10^{-17} c.c./sec. when the O_2 molecules formed are in the ground state. How large a factor should be included to allow for the production of vibrationally excited molecules is unknown. Information on this matter is urgently required for ionospheric theory. No data are available on the rate of (28*a*).

Detachment by collision with excited particles may be comparatively rapid—its coefficient has not yet been determined, but a value of order 10^{-10} c.c./sec. is suggested (Massey 1938).

5·4. *Transfer reactions*

The possible transfer reactions are:

$$O^+ + O_2 \rightarrow O + O_2^+ + 1{\cdot}3\,\text{eV}, \tag{30}$$

$$N_2^+ + O \rightarrow N_2 + O^+ + 2{\cdot}0\,\text{eV}, \tag{31}$$

$$N_2^+ + O_2 \rightarrow N_2 + O_2^+ + 3{\cdot}3\,\text{eV}, \tag{32}$$

$$O_2^- + O \rightarrow O_2 + O^- + 1{\cdot}2\,\text{eV}. \tag{33}$$

Under upper atmospheric conditions these are of course irreversible. It is not yet possible to give any definite values for the rates at which they occur. Some estimate might be made from a study of the passage of O^+, N_2^+ and O^- ions through O_2 and of O^+ and O_2^+ ions through N_2.

6. EQUILIBRIUM OF ATMOSPHERIC IONIZATION

In applying the information summarized in the preceding sections it is convenient to begin by considering the behaviour of a region with a single active constituent, the other constituents being passive except in contributing to the rate of three-body processes. This simplification shows clearly the relative importance of the various reactions and provides a wide basis to which modifications and extensions can be added.

If, at a particular level, n, n_e, n^+, n^- are the concentrations in numbers/c.c. of neutral particles, electrons, positive ions and negative ions respectively, and T is the absolute temperature, then, neglecting diffusion,

$$\frac{dn_e}{dt} = \mu n + \rho n^- + \kappa n^- n - \eta n_e n - \alpha_e n^+ n_e - \frac{n_e}{T}\frac{dT}{dt}, \tag{34}$$

$$\frac{dn^-}{dt} = \eta n_e n - \rho n^- - \kappa n^- n - \alpha_i n^+ n^- - \frac{n^-}{T}\frac{dT}{dt}, \tag{35}$$

$$n_e + n^- = n^+. \tag*{(36)*}$$

In these equations μn is the rate of the photo-ionization (§ 4), ρn^- the rate of photodetachment (§ 5·3), $\kappa n^- n$ the rate of collision detachment (§ 5·3), $\eta n_e n$ the rate of attachment (§ 5·2), $\alpha_e n^+ n_e$ the rate of electronic recombination (§ 5·1·1) and $\alpha_i n^+ n^-$ the rate of ionic recombination (§ 5·1·2). The terms involving T arise from the change of density with temperature. With

$$n^- = \lambda n_e, \tag{37}$$

then

$$\frac{dn_e}{dt} = \frac{\mu n}{1+\lambda} - \alpha n_e^2, \tag{38}$$

where

$$\alpha = \alpha_e + \lambda\alpha_i + \frac{1}{n_e T}\frac{dT}{dt} + \frac{1}{n_e(1+\lambda)}\frac{d\lambda}{dt}, \tag{39}$$

$$\lambda = \left\{\alpha_e n_e^2 + \frac{n_e}{T}\frac{dT}{dt} + \frac{dn_e}{dt} + \eta n_e n - \mu n\right\} \Big/ \{\kappa n_e n + \rho n_e - \alpha_e n_e^2\}. \tag{40}$$

---

* The electrostatic forces do not allow an appreciable departure from neutrality—a preponderance of positive ions (such as proposed by Ta-You Wu (1945)) is not possible.

α is the effective recombination coefficient, and $\mu n/(1+\lambda)$ the effective rate of electron production (q).

Ta-You Wu in a recent paper (1945) has proposed that the equations can be written approximately in the form

$$\frac{dn_e}{dt} = \mu n - \alpha_e(1+\lambda)\, n_e^2 - \frac{n_e}{T}\frac{dT}{dt}.$$

It is clear, however, that this is quite unjustified as it involves neglect of the major terms.

6·1. *The E layer*

For the E layer several simplifications can at once be made to the fundamental formulae. Not only is it unlikely that α_e is appreciable,* but in addition the terms dT/dt and $d\lambda/dt$ are small and can be neglected (except perhaps at sunrise and sunset) so that

$$\alpha = \lambda\alpha_i, \tag{41}$$

$$\lambda = \eta n_e n/(\kappa n_e n + \rho n_e + \alpha_i n_e^2 + q). \tag{42}$$

It will be noted that for the effective recombination coefficient to be pressure independent (as is observed) it is necessary in general that

$$\kappa n_e n \gg \rho n_e, \quad \alpha_i n_e^2 \quad \text{and} \quad q. \tag{43}$$

This very restrictive condition is of great importance in the following discussion. It will be noticed that the predominance of collision over photo-detachment ensures that the value of λ does not increase during the night. For the further discussion of the E layer we must consider three separate possibilities.

(i) The E layer below the atmospheric transition region—molecular oxygen being the only active constituent. With an attachment energy to O_2 of 1 eV (§ 5·1·2), no collision detachment process is available (§ 5·3).† Condition (43) cannot, therefore, be fulfilled and no satisfactory theory is possible. This conclusion suggests that the E layer is not formed below the atmospheric transition region and so enhances the density difficulty associated with the view that it arises from the ionization of O_2.

(ii) The E layer above the atmospheric transition region—atomic oxygen being the only active constituent. The consequences of condition (43) are again very important. Quantitative discussion is possible because of the data available concerning the rates of the various processes associated with atomic oxygen. q is only 220/c.c. sec. (§ 3·4). For photo-detachment (25*a*), ρ is most probably about 0·35/sec. (§ 5·3). The value of α_i for ionic recombination ((19), (20)) is uncertain but is unlikely to be greater than 1×10^{-7} c.c./sec. Thus, taking n_e as $1{\cdot}5 \times 10^5$ c.c.,

$$\rho n_e = 5{\cdot}2 \times 10^4/\text{c.c./sec.},$$

$$\alpha_i n_e^2 < 2{\cdot}2 \times 10^3/\text{c.c./sec.}$$

---

* The possibility that the rate of dissociative recombination (16) is grossly underestimated cannot be completely excluded but is not considered in the present paper. It is sufficient to mention that if the process is sufficiently rapid many difficulties disappear.

† Excited atoms are too rare for (29*b*) to be effective.

In view of (43) $\kappa n_e n$ must therefore be at least of order 2×10^5/c.c./sec. so that with n about 2×10^{12}/c.c.,

$$\kappa = 7 \times 10^{-13}\ \text{c.c./sec.}$$

Such a coefficient is far larger than estimated (§ 5·3) but cannot be regarded as impossible. Accepting it provisionally and using the result that for photo-attachment (§ 5·2) the most probable value of η is about $1{\cdot}1 \times 10^{-15}$ c.c./sec. λ can be seen from equation (42) to be only $1{\cdot}3 \times 10^{-3}$. Further, since the observed value of α is 1×10^{-8} c.c./sec., equation (41) gives α_i to be 8×10^{-6} c.c./sec. This value is much greater than predicted by theory (§ 5·1·2). As will be noted also it causes the term $\alpha_i n_e^2$, in the expression for λ, to become appreciable, and a recalculation making allowance for this produces a worsening of the position.

The differences between the derived and theoretical values of κ and α_i can be reduced by certain feasible changes in the numerical data used. Resonance may profoundly affect the photo-attachment and photo-detachment rates—η becoming $3{\cdot}0 \times 10^{-14}$ c.c./sec. at 500° K and ρ 0·85/sec., so that

$$\kappa = 1{\cdot}7 \times 10^{-12}\ \text{c.c./sec.}, \quad \alpha_i = 7 \times 10^{-7}\ \text{c.c./sec.}, \quad \lambda = 1{\cdot}4 \times 10^{-2}.$$

Thus though the discrepancy in connexion with the value of κ is slightly increased, that in connexion with the value of α_i is greatly reduced.*

A further improvement is obtained if, instead of taking the electron affinity of oxygen as 2·2 eV, the recent determination by Vier & Mayer (1944) is accepted, and it is taken as 3·1 eV (see §§ 5·2 and 5·3). This has the effect of multiplying the values of η and ρ by 2·8 and 0·4 respectively, with the result that with normal (non-resonance) photo-attachment and detachment we would now have

$$\kappa = 3 \times 10^{-13}\ \text{c.c./sec.}, \quad \alpha_i = 1{\cdot}4 \times 10^{-6}\ \text{c.c./sec.}, \quad \lambda = 7 \times 10^{-3},$$

and with resonant photo-attachment and detachment

$$\kappa = 6 \times 10^{-13}\ \text{c.c./sec.}, \quad \alpha_i = 1{\cdot}1 \times 10^{-7}\ \text{c.c./sec.}, \quad \lambda = 9 \times 10^{-2}.$$

Finally, the model used for the E layer may be inaccurate. The greatest uncertainty lies in the value taken for n. There is the possibility that this has been under-estimated and consequently that the value of κ and α_i required has been over-estimated. It should be noted, however, that in discussing the production of a layer by the ionization of atomic oxygen a smaller rather than a greater gas density appeared necessary (§ 4·5).

(iii) The E layer within the atmospheric transition region. The presence of both molecular and atomic oxygen in the E layer does not greatly affect the arguments advanced in (i) and (ii), but consideration has to be given to several additional processes. In particular, between O_2^- ions and O atoms either associative detachment,

$$O_2^- + O \rightarrow O_3 + e, \tag{44}$$

or electron transfer,

$$O_2^- + O \rightarrow O_2 + O^-, \tag{45}$$

---

* In certain circumstances di-electronic attachment and detachment ((22), (26)) can become very large and produce similar changes.

can occur.* Of these the latter is probably the more rapid. It is of interest because of the possibility that the stabilization involved might lead to a larger value of λ. This could occur if the various reaction rates were such as to make the main attachment be to O_2 and the main detachment be from O^-. A non-constant λ (and hence α) would, however, result because of the great increase with height of the amount of dissociation. As this is in conflict with observation it is to be assumed that the whole effect is unimportant. No other complications likely to influence the equilibrium are apparent. It would thus seem that the principal reactions for this possibility would simply be

$$O_2+h\nu\rightarrow O_2^+ +e, \tag{46}$$

or

$$O+h\nu\rightarrow O^+ +e, \tag{47}$$

followed by

$$O^+ +O_2\rightarrow O+O_2^+, \tag{48}$$

$$O+e\rightarrow O^- +h\nu. \tag{49}$$

$$O^- +O\rightarrow O_2+e, \tag{50}$$

$$\text{(and, less rapid, } O^- +h\nu\rightarrow O+e), \tag{51}$$

$$O^- +O_2^+\rightarrow O'+O_2'', \tag{52}$$

and perhaps

$$O^- +O^+\rightarrow O'+O''. \tag{53}$$

The quantitative discussion already given regarding the magnitudes κ, α_i and λ is quite unaffected.

Special mention should be made of the difficulty associated with the production of a layer within the transition region—if it originated from the ionization of O_2 it would be expected to occur at a lower level and if it originated from the ionization of O it would be expected to occur at a greater level (§ 4·5).

6·2. *The F_1 layer*

In some respects this layer should be comparatively easy to deal with theoretically, as atomic oxygen is the only constituent which can form negative ions. The discussion follows the same lines as for the *E* layer in which atomic oxygen was assumed as the only active constituent (6·1 (ii)). Even if molecular nitrogen plays a part in producing photo-ions at these levels this introduced no important new features—the additional reactions

$$N_2+h\nu\rightarrow N_2^+ +e, \tag{54}$$

$$N_2^+ +O\rightarrow N_2+O^+, \tag{55}$$

$$O^- +N_2^+\rightarrow O'+N_2'', \tag{56}$$

do not affect the general argument.

---

* The corresponding interactions between O^- ions and O_2 molecules are forbidden on energetic grounds.

Insertion of numerical values soon shows, however, that the theory of recombination in the F_1 layer encounters the same difficulties as for the E layer in a more exaggerated form. This is because the smaller effective recombination coefficient is more than compensated by the decreased density of atomic oxygen and the increased density of the ionization. Taking n as 3×10^{10}/c.c., n_e as $2{\cdot}5 \times 10^5$/c.c. and α as 4×10^{-9} c.c./sec. and assuming even resonant photo-attachment and detachment with the higher value of the electron affinity so that η (at 1000° K) is $4{\cdot}2 \times 10^{-14}$ c.c./sec. (§ 5·2) and ρ 0·34/sec. (§ 5·3), it can be seen that equation (42) gives such a small value for λ (7×10^{-4}) that equation (41) forces α_i to be so large (6×10^{-6} c.c./sec.) that

$$\alpha_i n_e^2 > \kappa n_e n, \quad \rho n_e. \tag{57}$$

This would imply physically that the main loss of negative ions is due, not to detachment, but to recombination. Electrons would disappear as if attachment were the only process involved. This would encounter the well-known difficulty of understanding why a layer would be formed at all as the effective recombination coefficient for electrons would be proportional to the pressure and the equilibrium electron density $(\mu n/\alpha)^{\frac{1}{2}}$, would tend to a constant at great heights. Furthermore, the required ionic recombination coefficient (α_i) of 6×10^{-6} c.c./sec. is several times larger than the value of $2{\cdot}4 \times 10^{-6}$ c.c./sec. found in air at atmospheric pressure by Sayers (1938) and is clearly unacceptable.

Relatively small changes in the values of α and n would improve the position as regards the F_1 layer. Thus taking n to be 6×10^{10}/c.c. and α as 2×10^{-9} c.c./sec. in F_1 (instead of the previously assumed 3×10^{10}/c.c. and 4×10^{-9} c.c./sec. respectively) and substituting as before shows that the undesirable predominance of the term $\alpha_i n_e^2$ is avoided. Further, if κ is as large as 2×10^{-11} c.c./sec. the quadratic law is obeyed and the effective recombination coefficient is constant—the necessary value of α_i being $1{\cdot}5 \times 10^{-6}$ c.c./sec. and the resultant value of λ, $1{\cdot}3 \times 10^{-3}$. This value of α_i is barely acceptable but, on substitution of the new value of κ in the analysis of the E region instead of the somewhat smaller value used in § 6·1 (ii), it is found that α_i there would have to be at least 3×10^{-6} c.c./sec. to provide the observed value of α. It would seem then that to provide a consistent picture of both the E and F_1 layers the rate of ionic recombination would have to be stretched beyond its possible limit. It is probable that the value required of the dissociative detachment coefficient is also too large but there is less definite evidence on this score. Before discussing the question further we will briefly consider the F_2 layer.

6·3. *The F_2 layer*

The complex and irregular nature of the F_2 layer arises from many causes. No attempt is made here to give a detailed theory including the influence of such effects as the temperature changes. Instead the discussion is limited to a short account of the fundamental processes.

The production of the layer is presumably due to the ionization of molecular nitrogen and atomic oxygen (§ 4·5). It is by no means improbable that the relatively slow recombination rate observed (α being 8×10^{-11} c.c./sec. by day and 3×10^{-10} c.c./sec. by night), arises at least partially from some form of electronic recombination. The most promising action is dissociative recombination

$$N_2^+ + e \rightarrow N' + N''. \tag{58}$$

It is uncertain however if the transfer collisions

$$N_2^+ + O \rightarrow N_2 + O^+ \tag{59}$$

are sufficiently rare to allow the required N_2^+ ions to be present throughout the night (§ 7·2·1·2). No recombination process to O^+ has yet been suggested that is at all rapid but it is unwise to exclude such a possibility completely until the anomalous position arising from the discharge tube measurements of Kenty and Mohler (§ 5·1·1) has been elucidated.

Ionic recombination can be discussed as for the E and F_1 layers. The conditions to be satisfied are less stringent in that there is no evidence that the effective recombination coefficient is pressure independent and in consequence there is no need for collision detachment to predominate. If the same values of the various coefficients are used as before but allowance is made for the probable higher temperature (2000° K) then on substitution in (40) it is seen that

$$\lambda = 8 \times 10^{-5} \quad \text{(day)}, \qquad 3 \times 10^{-4} \quad \text{(night)},$$

and

$$\alpha = 1{\cdot}1 \times 10^{-10} \quad \text{(day)}, \qquad 4{\cdot}5 \times 10^{-10} \quad \text{(night)}.$$

The agreement with the observed values of α is satisfactory. It is interesting to observe that owing to the magnitude of the term $\alpha_i n_e^2$ the negative ions disappear by recombination faster than by detachment so that

$$dn_e/dt \simeq \eta n_e n. \tag{60}$$

Mohler (1940) has found evidence in favour of such a law from the variation of the effective recombination coefficient ($\eta n/n_e$) during part of a sunspot cycle.

While in this way an adequate description of the layer can be obtained, it is worth noting that the very large coefficients required for the E and F_1 layers are not essential. Thus taking

$$\kappa = 2 \times 10^{-12} \text{ c.c./sec.} \quad \text{and} \quad \alpha_i = 1{\cdot}5 \times 10^{-7} \text{ c.c./sec.}$$

(values 10 times smaller than before) it can be seen that

$$\lambda = 3 \times 10^{-4} \quad \text{(day)}, \qquad 3 \times 10^{-3} \quad \text{(night)},$$

and

$$\alpha = 4{\cdot}4 \times 10^{-11} \text{ c.c./sec.} \quad \text{(day)}, \qquad 4{\cdot}5 \times 10^{-10} \text{ c.c./sec.} \quad \text{(night)},$$

in acceptable accord with observation.

6·4. *Summary and general discussion*

Provided the E layer lies at a height at which atomic oxygen occurs in adequate quantity a possible theory can be developed, though unexpectedly large collision detachment and ionic recombination coefficients are necessary. The F_1 layer can only be explained if still greater values of these constants be used. This has the effect of rendering the whole theory unconvincing and suggests that some factor such as the influence of atmospheric dust has been unjustifiably neglected. The essential difficulty lies in explaining the magnitude and the pressure independence of the observed effective recombination coefficients. Electronic recombination would appear to be quite inadequate, and if the assumed atmospheric density is correct, the ionic recombination rate could hardly be fast enough to compensate for the very low value of the ratio λ of negative ion to electron density which must result from the occurrence of photo-detachment. Moreover, the collision detachment is unlikely to be sufficiently rapid to predominate as would be required to produce an effective recombination coefficient independent of pressure.

The smallness of the important ratio λ would seem inescapable. It does not conflict with information from other sources. No reliable direct radio evidence has yet been obtained. At one time, however, it was thought that in the E layer a large ratio (of order 100) was required to provide sufficient conductivity for the circulating currents which, according to the dynamo theory, are responsible for the magnetic variations associated with the atmospheric tides. An unknown factor which entered into the calculations was the amplitude of the tidal oscillations in the upper atmosphere as compared with those near ground level. A sufficient amplification would render the large ratio unnecessary. Pekeris (1937) showed that for at least the solar tide, the amplification is, in fact, probably considerable due to resonance. However, a complete review of the whole problem has now been rendered essential by the observation by Appleton & Weeks (1939) of variations in the height of the E layer correlated with the lunar tides with amplitude 7×10^3 that at ground level and with phase opposite to that required by the dynamo theory. Until this remarkable effect is explained it is clearly impossible to draw any conclusions on the magnitude of the negative ion electron ratio from the magnetic variations.

The position regarding the F_2 layer is rather more satisfactory than that for the lower layers there being less difference between observation and the predicted rates of either electronic or ionic recombination. Even here, however, the quantitative information is inadequate to establish fully the processes involved.

While there are, as has been insisted, grave difficulties in the present theory of the ionized layers it is perhaps desirable for clarity to recapitulate in general terms the description given. The basic reactions determining the behaviour are as follows:

$$\text{Photoionization:}\quad h\nu + O_2 \rightarrow O_2^+ + e,\quad h\nu + O \rightarrow O^+ + e,\quad h\nu + N_2 \rightarrow N_2^+ + e. \tag{61}$$

$$\text{Charge transfer:}\quad N_2^+ + O \rightarrow O^+ + N_2,\quad O^+ + O_2 \rightarrow O_2^+ + O. \tag{62}$$

$$\text{Radiative attachment:}\quad O + e \rightarrow O^- + h\nu. \tag{63}$$

Associative detachment: $O+O^- \rightarrow O_2+e.$ (64)

Photo-detachment: $h\nu+O^- \rightarrow O+e.$ (65)

Ionic recombination:

$$O^-+O_2^+ \rightarrow O'+O_2', \quad O^-+O^+ \rightarrow O'+O'', \quad O^-+N_2 \rightarrow O'+N_2'. \tag{66}$$

Electronic recombination: $N_2^+ + e \rightarrow N'+N''.$ (67)

The electrons freed by the photo-ionization attach themselves by a radiative process (63) to the neutral oxygen atoms to form negative ions. These can be destroyed by associative detachment (64) or by photo-detachment (65) to yield electrons once more or by ionic recombination (66) to yield uncharged particles. For the E and F_1 layers it is necessary for the first of these processes to be the most rapid in order to explain the pressure independence of the effective recombination coefficient. The difference in the effective recombination coefficients in the two layers can be ascribed partially to temperature effects and partially to the nature of the positive ions involved (which depend not only on the particular photo-ionization process (§ 61) but possibly in addition on charge transfer (62)). For the F_2 layer, in contrast to the E and F_1 layers, the associative detachment process is thought to be unimportant because of the low pressure so that ionic recombination (66) gives a pressure dependent effect. It is not unlikely also that electronic recombination (67) is appreciable —especially during the day.

Plausible theories of the layers have been put forward by other workers—notably Tukada (1937) and Mohler (1940). It must be emphasized however that they are unsound and are apparently successful only because they ignore the full effects of photo-detachment. No processes beyond those discussed in this paper are taken into account.

7. RADIATION OF THE NIGHT SKY (NON-POLAR AURORA)

Before attempting any correlation of night-sky radiation phenomena with the foregoing considerations it is desirable to summarize briefly the observed data. Exhaustive references are not given.

7·1. *Observed data**

The chief contributors to the night-sky emission spectrum are neutral atoms of oxygen and sodium and neutral molecules of nitrogen. It is convenient to collect the available information under the heading of the atomic or molecular source concerned.

7·1·1. *Radiation from atomic oxygen*

The two forbidden transitions ^1D–^1S and ^3P–^1D give rise to the auroral green (λ5577 A) and nebular red (λ6300 A, λ6364 A) lines. The average intensity corresponds to the emission of 2×10^8 quanta/cm.^{2} column of atm./sec. (Rayleigh 1930; Cerniajev,

---

* See Dejardin (1936), Elvey (1942), and Pearse (1943).

Khvostikov & Panschin 1936) for the green line and about 1×10^7 quanta/cm.^{2} column of atm./sec. (estimated) for the red. Condon (1934) has calculated the relevant radiative life-times of the ^1S and ^1D states of O to be 0·5 and 100 sec. respectively. It follows that the number of atoms/cm.^{2} column of atmosphere in the ^1S state is of order 1×10^8, and in the ^1D state 1×10^9.

Information concerning the diurnal, seasonal and other variations in intensity of the lines is slowly accumulating. McLennan (1928) and other writers found the intensity of λ 5577 A to have a pronounced maximum about midnight. The regular occurrence of this effect is not yet established. Instead Elvey & Farnsworth (1942) found little or no such maximum, while Karandikar (1934) observed a midnight minimum. The factors determining which type of intensity variation occurs are not yet clear. The intensity of λ 6300 A has been observed by Garrigue (1936) and by Elvey, Swings & Linke (1941) to decrease from sunset to about midnight and then to remain practically constant.

No appreciable twilight effects have been found for λ 5577 A. In contrast λ 6300 A exhibits considerable enhancement—the intensity when the sunlight strikes the atmosphere above 70 km. being of order 10 times that during the night (Elvey & Farnsworth 1942). A remarkable feature of the phenomenon is the appreciable magnitude of the enhancement even when the sunlight can only reach the very high atmosphere.

Measurements of Rayleigh & Jones (1935) at Terling (England), Capetown (South Africa) and Canberra (Australia) show that there are large irregular local fluctuations of intensity of the green and red region of the spectrum on which are superimposed in addition a semi-annual and annual variation and a slow secular change with sunspot cycle—an interesting similarity to the frequency of occurrence of aurorae. The effects depend to some extent on the station.

Several attempts have been made to measure the height of the source of the radiations. The most recent work is that of Elvey & Farnsworth (1942) who made 25 determinations of the variation of intensity with zenith distance at a number of stations in North and South America for λ 5577 A and λ 6300 A. The average variation was much the same for both lines and indicated the great height of 500 km. for the emitting layer. This result is in general agreement with the earlier investigations of Cabannes & Dufay (1934), of Brunner (1935) and of Garrigue (1936). It is so important however that more measurements are required to confirm it.

7·1·2. *Radiation from atomic sodium*

The presence of the D lines (λ 5893 A, ^2S–^2P) of neutral sodium in the night-sky spectrum has been established by Cabannes, Dufay & Gauzit (1938). The number of quanta emitted in these lines is about 1×10^7/cm.^{2} column of atm./sec. Apart from fluctuations the intensity remains fairly constant throughout the night. There is a very strong twilight effect, the enhancement being of the order 100 when the atmosphere above 70 km. is illuminated. Measurements by Bernard (1938), and Elvey & Farnsworth (1942), on the variation of the intensity of the twilight radiation with zenith distance show that at least between 70 and 115 km. the concentration of emitting atoms falls off at much the same rate as that of the main gas.

Systematic study of the intensity variations exhibited by the D lines has hardly begun but the emission appears to be least in summer (Garrigue (1937), Elvey *et al.* (1941)).

Cabannes *et al.* (1938) estimate that the height of layer responsible for the lines is 130 km.

7·1·3. *Radiation from molecular nitrogen*

Most of the night-sky band emission* is due to neutral nitrogen. The strongest are the bands of the Vegard-Kaplan system arising from optically forbidden transitions between the $A\,^3\Sigma$ and the ground $X\,^1\Sigma$ electronic states and having excitation energy 7 eV. Following in intensity are the bands of the Lyman-Birge-Hopfield system ($A\,^1\Pi \rightarrow X\,^1\Sigma$) of excitation energy 8·5 eV and those of the First Positive system ($B\,^3\Pi \rightarrow A\,^3\Sigma$) of excitation energy 10 eV. The Second Positive system ($C\,^3\Pi \rightarrow B\,^3\Sigma$) has not been observed to be present. Reliable information on the intensity of the bands is not available but as a very rough estimate it may be taken that the number of emitted quanta/cm.^{2} column of atm./sec. is 1×10^7 for the Vegard-Kaplan system, 6×10^6 for the Lyman-Birge-Hopfield system and 3×10^6 for the First Positive system. No information is available as to the radiative lifetime of the metastable $A\,^3\Sigma$ state but it must be long.

Slipher (1933) observed in the twilight spectrum a flash due to the emission of the Negative band system ($B\,^2\Sigma \rightarrow X\,^2\Sigma$) of N_2^+ (excitation energy 18·7 eV).

The variations in the blue region of the spectrum to which the N_2 bands largely contribute have been studied by several workers. Rayleigh (1929) and Bradbury & Sumerlin (1940) observed that the intensity has a maximum at midnight. Rayleigh & Jones (1935) found fluctuations in close correlation with those of the oxygen green and red lines and also corresponding semi-annual and annual variations and slow secular change with sunspot cycle.

7·1·4. *Other night-sky radiations*

As well as the main radiations described above, there is evidence of the emission of the Schumann-Runge bands ($B\,^3\Sigma \rightarrow X\,^3\Sigma$) of O_2, and possibly of the β ($B\,^2\Pi \rightarrow X\,^2\Pi$) and γ ($A\,^2\Sigma \rightarrow X\,^2\Pi$) bands of NO. A number of bands, particularly a strong one at λ 3556 A, remain unidentified. There is no radiation which can definitely be attributed to atomic nitrogen, atomic or molecular hydrogen, or any of the inert gases.

7·1·5. *The spectrum of the polar aurora*

Apart from the well-known high degree of excitation as indicated by the lines of neutral and ionized atomic oxygen and nitrogen and by the relative intensities of the nitrogen band systems (in particular the great strength of the Negative system) the main feature of interest in the spectrum of the polar aurora as compared with that of the night sky is the absence, or at least the comparative faintness, of the sodium lines, and the oxygen bands.

---

* The identification of bands is very difficult and is still open to discussion. In this paper we have accepted in general the suggestions of Elvey *et al.* (1941).

7·2. *Origin of night-sky emission*

We now come to the difficult problem of the origin of the excitation which gives rise to the radiation described. We treat separately the nocturnal and the twilight emission.

7·2·1. *The nocturnal emission*

7·2·1·1. *The oxygen recombination theory.* A possible source of the energy needed for the emission lies in the region of dissociated oxygen. Chapman (1931) suggested that when oxygen atoms recombine to form a molecule excitation energy may be given to some third body.

For the green and red lines the basic reactions are

$$\mathrm{O}+\mathrm{O}+\mathrm{O}\rightarrow\mathrm{O}_2+\mathrm{O}'\begin{cases}(^1\mathrm{S}) & (68)\\ (^1\mathrm{D}), & (69)\end{cases}$$

and for the D lines,

$$\mathrm{O}+\mathrm{O}+\mathrm{Na}\rightarrow\mathrm{O}_2+\mathrm{Na}'(^2\mathrm{P}). \qquad (70)$$

However, the energy required to excite molecular nitrogen for the emission of the Vegard-Kaplan, Lyman-Birge-Hopfield, and First Positive band systems is greater than the dissociation energy of oxygen. As the particles concerned in the collision processes are in thermal equilibrium, the additional energy cannot come from their relative motion. However, if one of the recombining oxygen atoms is in the metastable $^1\mathrm{S}$ or $^1\mathrm{D}$ states sufficient energy is available at least for the Vegard-Kaplan and Lyman-Birge-Hopfield band systems. Thus,

$$\mathrm{O}'\begin{cases}(^1\mathrm{S}) \\ (^1\mathrm{D})\end{cases}+\mathrm{O}+\mathrm{N}_2\rightarrow\mathrm{O}_2+\mathrm{N}_2'(A\,^3\Sigma) \quad \text{(giving the Vegard-Kaplan band system),} \qquad (71),\ (72)$$

$$\mathrm{O}'(^1\mathrm{S})+\mathrm{O}+\mathrm{N}_2\rightarrow\mathrm{O}_2+\mathrm{N}_2'(A\,^1\Pi) \quad \text{(giving the Lyman-Birge-Hopfield band system).} \qquad (73)$$

At first sight the theory is attractive, but on closer examination certain difficulties become manifest particularly for the sodium and nitrogen reactions.

As far as the green and red lines are concerned the intensity likely to result may well be adequate—the total number of oxygen recombinations must be of the order 10^{12}/cm.^{2} column of atm./sec. and it is quite possible that a sufficiently large fraction of the recombining atoms dispose of their excess energy in the manner suggested. On the other hand if t is the lifetime of a metastable oxygen atom towards deactivation by collision, and τ is the radiative lifetime, the number of photons emitted at any level by metastable oxygen atoms produced by (68), (69) is proportional to

$$p^3 t/(t+\tau), \qquad (74)$$

where p is the partial pressure of atomic oxygen. This has a maximum at the level of maximum density of atomic oxygen—that is just over 100 km., in contradiction to the results described in (§ 7·1·1) which suggests a much higher level for the origin of the lines.* A satisfactory theory must also include an explanation of the intensity variations. As the rate of recombination of oxygen atoms is so slow that there is a negligible decrease in their concentration during the night, little diurnal variation in intensity would be expected. The observational evidence on this is still incomplete but it certainly indicates a difference in the behaviour of the green and red lines that would not have been predicted. Further the fluctuations of intensity are not easy to understand. Little can be said at this stage about the possibility of explaining the semi-annual and annual variations, the slow secular change with sunspot cycle, and the dependence on station.

It thus appears that the application of the theory to the green and red lines is not free from serious difficulties. It must be noted that these depend on experimental results which are as yet far from definite. The position is far worse when the excitation of the sodium lines and nitrogen bands is considered.

The ratio of the numbers of sodium and oxygen atoms in the upper atmosphere is probably about $5\text{–}50 \times 10^{-11}$ ($5\text{–}50 \times 10^{8}/10^{19}$) (§§ 2·3, 7·2·2) yet the intensity of emission of the D lines is as great as 5×10^{-2} of that of the green lines. As the rates of the reactions (68), (69) and (70) are not likely to differ greatly there is thus a very grave discrepancy even allowing for a considerable rate of collision de-activation of the metastable oxygen atoms.

A similar position is encountered when the effectiveness of the reactions (71), (72) and (73) in providing excited nitrogen molecules is estimated. In the nocturnal upper atmosphere the fraction of oxygen atoms in the metastable ^1S and ^1D states is only 10^{-11}, ($10^{8}/10^{19}$), and 10^{-10}, ($10^{9}/10^{19}$), respectively. The intensities of the nitrogen band systems would therefore be very many orders of magnitude lower than those of the oxygen lines—in complete disagreement with observation.

It seems necessary to seek explanations of the excitations leading to the D lines and the nitrogen band systems other than through Chapman's theory.

7·2·1·2. *The ionized layer theory.* It is natural to consider that there may be some connexion between the nocturnal emission and the ionized layers. However Chapman (1931) has shown that the rate of emission allowable by the ionization density is likely to be too low for some of the radiations.

The green line is maintained at a rate of 2×10^{8} quanta/cm.^{2} column of atm./sec. so that about 10^{13} quanta/cm.^{2} column of atmosphere are emitted during the night. If the emission is limited to a normal layer of scale height H the total number of quanta (at the height of greatest emission) must be about $10^{13}/2{\cdot}7\,H$/cm.^{2}, which for the E layer is 4×10^{6}/c.c. ($H = 1 \times 10^{6}$ cm.) and for the F layer is 9×10^{5}/c.c. ($H = 4 \times 10^{6}$ cm.). To account for the constancy of the emission and to allow for an efficiency of less than unity for the process involved a maximum ion or electron density of some 4×10^7/c.c. is required in the E layer and of some 9×10^6/c.c. in the F layer.* These values are much higher than those observed, particularly for the E region. The production of the less intense red lines does not lead to the same difficulty, at least as regards the F region, and it is worth considering the possible processes in more detail.

---

* It can be generalized that no two- or three-body collision processes can give rise mainly to emission from a layer high in the atmosphere unless one of the particles involved is localized there.

The number of excited particles/cm.^{2} column of atm./sec. produced by electronic and ionic recombination is $2{\cdot}7H(1+\lambda)\,n_e^2\alpha_e$ and $2{\cdot}7H\lambda(1+\lambda)\,n_e^2\alpha_i$ respectively. λ is small. The night-time value of n_e in the E layer is, as we have seen in § 3·4, about 10^4 and in the F layer is about $2{\cdot}5 \times 10^5$. For radiative or di-electronic recombination

$$\mathrm{O} + e \rightarrow \mathrm{O}'({}^1D) + h\nu, \tag{75}$$

$$\left.\begin{aligned} &\mathrm{O}^+ + e \rightarrow \mathrm{O}'^{(D)} \quad \text{(doubly excited)},\\ &\mathrm{O}'^{(D)} \rightarrow \mathrm{O}'({}^1D_2) + h\nu. \end{aligned}\right\} \tag{76}$$

the recombination coefficient is less than 10^{-12} c.c./sec. so that the number of ^{1}D atoms formed/cm.^{2} column of atm./sec. is less than 3×10^2 for the E layer and 7×10^5 for the F. The dissociative recombination process

$$\mathrm{O}_2^+ + e \rightarrow \mathrm{O}'({}^1\mathrm{D}) + \mathrm{O}'', \tag{77}$$

may occur in the E layer but its recombination coefficient is unlikely to be great enough for the emission to be appreciable. For ionic recombination,

$$\mathrm{O}^+ + \mathrm{O}^- \rightarrow \mathrm{O}'({}^1\mathrm{D}) + \mathrm{O}'', \tag{78}$$

$$\mathrm{O}_2^+ + \mathrm{O}^- \rightarrow \mathrm{O}_2' + \mathrm{O}'({}^1\mathrm{D}), \tag{79}$$

$$\mathrm{N}_2^+ + \mathrm{O}^- \rightarrow \mathrm{N}_2' + \mathrm{O}'({}^1\mathrm{D}), \tag{80}$$

$$\mathrm{O}^+ + \mathrm{O}_2^- \rightarrow \mathrm{O}'({}^1\mathrm{D}) + \mathrm{O}_2', \tag{81}$$

the recombination coefficient is unknown but in certain cases it may be very large (§ 5·1·2). The interpretation of ionospheric recombination is still uncertain but it may indeed be that the value of $\lambda\alpha_i$ is of order 10^{-8} c.c./sec. in the E layer and 3×10^{-10} c.c./sec. in the F layer. The number of excited atoms formed/cm.^{2} column of atm./sec. in the E and F layers *may* thus be as great as 3×10^6 and 2×10^8 respectively. It seems quite possible therefore that ionic recombination in the F layer may give an important contribution to the intensity of the red lines, particularly during the early part of the night. A further source of ^{1}D atoms is from excitation by energetic electrons formed by dissociative detachment

$$\mathrm{O}^- + \mathrm{O} \rightarrow \mathrm{O}_2 + e \quad (+2{\cdot}9\,\mathrm{eV}\ \text{kinetic energy}), \tag{82}$$

$$e + \mathrm{O} \rightarrow \mathrm{O}'({}^1\mathrm{D}) + e. \tag{83}$$

---

* The argument is of course invalid if each ion or electron can produce the emission of a number of quanta. It is not easy to see how this could occur for the green line but for some of the other radiations the possibility exists ((82), (83)). Nocturnal renewal of ionization is also ignored.

A similar set of reactions was suggested by Martyn & Pulley (1936) to account for the green line but for this the electrons have insufficient energy.

It can be seen at once that the *D* lines cannot arise from any of the processes just discussed owing to the relative scarcity of sodium ions and atoms.

The nitrogen band systems, like the red lines, may well be produced by ionic recombination processes in the *F* layer. There may be some difficulty, however, owing to the transfer,

$$N_2^+ + O \rightarrow N_2 + O^+, \tag{84}$$

removing the nitrogen ions too rapidly—unless the probability/collision is very small ($<10^{-4}$) there will be no nitrogen ions left within a few hours of sunset.

If the ionized layer theory of the origin of the red lines and the nitrogen band systems is correct there should of course be some connexion between their intensity and the properties of the layer as determined by radio measurements,* but this may be very indirect owing to the nature of the many processes involved.

7·2·1·3. *The sodium oxide theory.* Neither the oxygen recombination theory nor the ionized layer theory seem able to explain the intensity of the *D* lines. In view of the comparative rarity of the sodium atoms it might be expected that some special process is required. Chapman (1939) suggested as a possible source of the excited atoms the action

$$NaO + O \rightarrow Na'(^2P) + O_2. \tag{85}$$

The formation of the necessary sodium oxide can proceed by such processes as,

$$Na + O + M \rightarrow NaO + M, \tag{86}$$

$$Na + O_3 \rightarrow NaO + O_2, \tag{87}$$

(and similarly for other suitable oxides).

None of these reactions is likely to oxidize sodium atoms at any considerable rate. The three-body reaction (86) is extremely slow at such heights as 100 km. Though (87) may be more rapid it is doubtful if even here the yield is adequate owing to the rarity of O_3. Assuming, however, that the concentration of NaO is not many orders smaller than that of Na the rate of (85) may well be sufficiently great *provided it is endothermic.* Since the dissociation energy of O_2 is 5·1 eV and the excitation energy of the ^2P state of Na is 2·1 eV this condition requires the dissociation energy of NaO to be less than 3·0 eV—if it is as great as 4 eV (as given by Chapman (1939)) some new theory for the origin of the D lines must be sought.†

7·2·1·4. *Conclusion.* It is clear that the origins of the night-sky emission are still far from certain. While none of the theories discussed are free from objection it is at least possible that the oxygen recombination theory accounts for the intensity of the green line and some of that of the red lines; the ionized layer theory accounts for the remainder of the intensity of the red lines and that of the nitrogen band systems; and the sodium oxide theory accounts for the intensity of the *D* lines.

---

* Mitra (1945) has recently reported evidence of a correlation between ion density in the *F* region and the intensity of night sky light emission when both are abnormal.

† The suggestion (Chapman 1939) that additional energy might be produced from reactions such as (85) if the oxygen atoms involved were in a metastable state cannot be accepted, as the number of such atoms is extremely small (§ 7·2·1·1).

If this picture is valid then:

(i) No correlation between the green line and nitrogen band intensity variations would be expected.

(ii) The height of maximum intensity of green light emission must not be far above the region of maximum density of atomic oxygen (just over 100 km.).

(iii) A considerable proportion of ionization in the *F* layer must remain in the form of N_2^+ throughout the night in spite of the transfer reaction (84).

(iv) The dissociation energy of NaO must be less than 3·0 eV.

The observational evidence available rather tends against both (i) and (ii) and the position must remain obscure until more extensive and reliable data are available.

7·2·2. *Twilight emission*

The most obvious source of twilight emission effects is from direct excitation by solar radiation. If only the resonance transition is involved the number of quanta emitted is simply

$$(\omega_E/\omega_N)\,\tau^{-1}\,e^{-h\nu/kT}f/\text{particle/sec.}^* \tag{88}$$

where ω_E and ω_N are the statistical weights of the excited and ground states respectively, τ the radiative life time, ν the frequency of the exciting radiation, T the absolute temperature of the solar black body (6000° K), and f the dilution factor ($5{\cdot}44\times10^{-6}$). This formula can readily be modified to apply to more complicated cases. For the green and red lines of oxygen and the D lines of sodium numerical values are given in table 5.

TABLE 5

emission line		ω_E/ω_N	τ (sec.)	transitions/atom/sec.
O	green	1/9	0·5	3×10^{-11}†
O	red	5/9	100	7×10^{-10}
Na	D	3	$1{\cdot}6\times10^{-8}$	20

Unfortunately the importance of collision deactivation of the metastable atoms is uncertain but (taking the number of oxygen atoms/cm.^{2} column of atmosphere to be 10^{19}) it would appear that direct excitation while not affecting the green line might enhance the red. Another possibility for this latter lies in the dissociation of molecular oxygen

$$O_2+h\nu\rightarrow O'(^1D)+O(^3P). \tag{89}$$

---

[*] It is implicitly assumed of course that the number of solar quanta of the correct wavelength is sufficient—this is correct even for the forbidden transitions on account of their Doppler broadening.

[†] Allowance has here to be made for the fact that the absorption is from the ^3P state but the emission is to a ^1D state.

It is doubtful however if this latter source would account for the observation by Elvey and Farnsworth of an enhancement at very great heights (§ 7·1·1). The process

$$O^- + h\nu \rightarrow O'(^1D) + e, \tag{90}$$

is perhaps of interest in this respect. From the calculations on photo-detachment (Bates & Massey 1943) it is found that the number of oxygen atoms raised to the ^1D state by reactions of the type (90) is from 2×10^{-2}/negative ion/sec. to 5×10^{-2}/negative ion/sec. (depending on the existence or not of resonance effects, § 5·2). Taking the negative ion density in the F region as 10^{-3} of that of the electrons and a layer thickness of 100 km. (both estimations being only tentative), this would give a rate of production of ^1D oxygen atoms of between 5×10^7 and 10^8/cm.^{2} column/sec., comparable with the nocturnal emission rate of about 10^7/cm.^{2} column/sec.

Direct excitation seems to be the only process likely to be important for the D lines. From above 70 km. the intensity of the line is about 1×10^9/cm.^{2} column of atm./sec. Taking into account the fact that Fraunhofer absorption reduces the solar radiation involved by a factor probably between 10 and 100* the number of sodium atoms/cm.^{2} column of atmosphere can be estimated to be between 5×10^8 and 5×10^9, as used in preceding sections.

REFERENCES

Appleton 1937 *Proc. Roy. Soc.* A, **162**, 451.
Appleton 1939*a* *Quart. J. R. Met. Soc.* **65**, 324.
Appleton 1939*b* *Occ. Notes, R. Astr. Soc.* **3**, 33.
Appleton & Chapman 1935 *Proc. Inst. Radio Engrs., N.Y.*, **23**, 658.
Appleton & Naismith 1935 *Proc. Roy. Soc.* A, **150**, 685.
Appleton & Naismith 1939 *Phil. Mag.* **27**, 144.
Appleton & Naismith 1940 *Proc. Phys. Soc.* **52**, 402.
Appleton, Naismith & Ingram 1937 *Phil. Trans*, A, **236**, 191.
Appleton & Weeks 1939 *Proc. Roy. Soc.* A, **171**, 171.
Bailey & Martyn 1935 *Wireless Engr.* **12**, 122.
Bates 1939 *Mon. Not. Roy. Astr. Soc.* **100**, 25.
Bates, Buckingham, Massey & Unwin 1939 *Proc. Roy. Soc.* A, **170**, 322.
Bates & Massey 1941 *Proc. Roy. Soc.* A, **177**, 329.
Bates & Massey 1943 *Phil. Trans.* A, **239**, 269.
Berkner 1939 *Terrestrial magnetism and electricity* (ed. Fleming). McGraw Hill.
Bernard 1938 *Z. Phys.* **110**, 291.
Best, Farmer & Ratcliffe 1938 *Proc. Roy. Soc.* A, **164**, 96.
Bhar 1938 *Indian J. Phys.* **12**, 363.
Bhar 1939 *Indian J. Phys.* **13**, 253.
Bloch & Bradbury 1935 *Phys. Rev.* **48**, 689.
Booker & Berkner 1938 *Terr. Magn. Atmos. Elect.* **43**, 427.
Bradbury 1933 *Phys. Rev.* **44**, 883.
Bradbury & Sumerlin 1940 *Terr. Magn. Atmos. Elect.* **45**, 19.
Brunner 1935 *Publ. Sternw. Polyt. Zürich*, **6**, 1.
Budden, Ratcliffe & Wilkes 1939 *Proc. Roy. Soc.* A, **171**, 188.
Cabannes & Dufay 1934 *C.R. Acad. Sci., Paris*, **198**, 306.

---

* A. Hunter (private communication, 7 October 1943).

Cabannes, Dufay & Gauzit 1938 *C.R. Acad. Sci., Paris,* **206,** 1525.
Cerniajev, Khvostikov & Panschin 1936 *J. Phys. Radium,* **7,** 149.
Chapman 1931 *Proc. Roy. Soc.* A, **132,** 353.
Chapman 1937 *Phil. Mag.* **23,** 657.
Chapman 1939 *Astrophys. J.* **90,** 309.
Chapman 1943 *Rep. Progr. Phys.* **9,** 92.
Chapman & Price 1936 *Rep. Progr. Phys.* **3,** 42.
Colwell & Friend 1936 *Nature,* **137,** 782.
Condon 1934 *Astrophys. J.* **79,** 217.
Dejardin 1936 *Rev. Mod. Phys.* **8,** 1.
Dellinger 1937 *Terr. Magn. Atmos. Elect.* **42,** 49.
Eckersley 1935 *Nature,* **135,** 435.
Eckersley 1939 *Nature,* **143,** 33.
Eckersley 1940 *Terr. Magn. Atmos. Elect.* **54,** 25.
Eckersley & Millington 1939 *Proc. Phys. Soc.* **51,** 110.
Elvey 1942 *Rev. Mod. Phys.* **12,** 140.
Elvey & Farnsworth 1942 *Astrophys. J.* **96,** 451.
Elvey, Swings & Linke 1941 *Astrophys. J.* **93,** 337.
Farmer & Ratcliffe 1935 *Proc. Roy. Soc.* A, **151,** 370.
Gardner 1938 *Phys. Rev.* **53,** 75.
Garrigue 1936 *C.R. Acad. Sci., Paris,* **202,** 1807.
Garrigue 1937 *C.R. Acad. Sci., Paris,* **205,** 491.
Gauzit 1943 *C.R. Acad. Sci., Paris,* **217,** 179.
Gilliland, Kirby, Smith & Reymer 1936 *Terr. Magn. Atmos. Elect.* **41,** 379.
Goodall 1939 *Proc. Inst. Radio Engrs, N.Y.,* **27,** 701.
Hagstrum & Tate 1941 *Phys. Rev.* **59,** 354.
Higgs 1942 *Mon. Not. R. Astr. Soc.* **102,** 24.
Hulburt 1939 *Phys. Rev.* **55,** 639.
Hunter 1943 *Rep. Progr. Phys.* **9,** 5.
Karandikar 1934 *Indian J. Phys.* **8,** 547.
Kenty 1928 *Phys. Rev.* **32,** 621.
Kirby, Berkner & Stuart 1934 *Proc. Inst. Radio Engrs, N.Y.,* **22,** 481.
Kirby, Gilliland & Judson 1936 *Proc. Inst. Radio Engrs, N.Y.,* **24,** 1027.
Loeb 1935 *Phys. Rev.* **48,** 684.
Lozier 1934 *Phys. Rev.* **46,** 268.
McLennan 1928 *Proc. Roy. Soc.* A, **120,** 327.
Maeda 1937 *Electrotech. J., Japan,* **1,** 66.
Martyn & Pulley 1936 *Proc. Roy. Soc.* A, **154,** 455.
Massey 1937 *Proc. Roy. Soc.* A, **163,** 542.
Massey 1938 *Negative Ions.* Camb. Univ. Press.
Massey & Bates 1943 *Rep. Progr. Phys.* **9,** 62.
Mitra 1945 *Nature,* **155,** 786.
Mitra & Bhar 1936 *Science and Culture,* **1,** 782.
Mitra, Bhar & Ghosh 1938 *Indian J. Phys.* **12,** 455.
Mohler 1937 *Bur. Stand. J. Res., Wash.,* **19,** 447, 559.
Mohler 1940 *Bur. Stand. J. Res., Wash.,* **25,** 507.
Pearse 1943 *Rep. Progr. Phys.* **9,** 42.
Pekeris 1937 *Proc. Roy. Soc.* A, **158,** 650.
Pierce 1941 *Phys. Rev.* **59,** 625.
Pierce, Higgs & Halliday 1940 *Phys. Rev.* **58,** 1119.
Price 1943 *Rep. Progr. Phys.* **9,** 10.
Rayleigh 1929 *Proc. Roy. Soc.* A, **124,** 395.
Rayleigh 1930 *Proc. Roy. Soc.* A, **129,** 458.
Rayleigh & Jones 1935 *Proc. Roy. Soc.* A, **151,** 22.
Saha & Rai 1938 *Proc. Nat. Inst. Sci., India,* **1,** 227.
Sayers 1938 *Proc. Roy. Soc.* A, **169,** 83.

Slipher 1933 *Mon. Not. R. Astr. Soc.* **93**, 657.
Sutherland & Callendar 1943 *Rep. Progr. Phys.* **9**, 18.
Ta-You Wu 1945 *Terr. Mag. Atmos. Elect.* **50**, 57.
Thomson 1924 *Phil. Mag.* **47**, 337.
Toshniwal, Pant & Bajpai 1936 *Nature*, **138**, 37.
Tukada 1937 *Rep. Radio Res., Japan*, **7**, 121.
Vier & Mayer 1944 *J. Chem. Phys.* **12**, 28.
Watson-Watt, Bainbridge-Bell, Wilkins & Bowen 1936 *Nature*, **137**, 866.
Wells & Stanton 1938*a* *Terr. Magn. Atmos. Elect.* **43**, 169.
Wells & Stanton 1938*b* *Terr. Magn. Atmos. Elect.* **43**, 467.
Wells & Stanton 1939 *Terr. Magn. Atmos. Elect.* **44**, 195.
Whipple 1938 *Proc. Amer. Phil. Soc.* **79**, 499.
Whipple 1943 *Rev. Mod. Phys.* **15**, 246.
White & Brown 1935 *Proc. Roy. Soc.* A, **153**, 639.
Wilkes 1939 *Proc. Phys. Soc.* **51**, 138.
Wulf & Deming 1938 *Terr. Magn. Atmos. Elect.* **43**, 283.
Yamanouchi & Katani 1940 *Proc. Phys.-Math. Soc., Japan*, **22**, 60.

The effect of an electric field on the viscosity of liquids

By E. N. da C. Andrade, F.R.S., and C. Dodd, Ph.D.

(*Received* 26 *November* 1945)

An apparatus for measuring the effect of an electric field on the viscosity of liquids has been developed, in which the liquid runs in a narrow channel between plane metal boundaries which can be used as electrodes. The apparatus allows the use of fields up to an average strength of 35 kV/cm. and gives a high degree of accuracy. With non-polar liquids, either dry or contaminated with water so as to conduct, and with non-conducting polar liquids an electric field transverse to the line of flow has no marked effect on the viscosity. With polar liquids that conduct relatively well, either normally or when contaminated with water, the electric field causes a considerable increase in apparent viscosity. This effect increases with increasing strength of field, but reaches a limiting value for very high fields. With polar liquids whose conductivity can be diminished by successive purifications the magnitude of this limiting effect is proportional to the conductivity of the specimen.

With an alternating field the effect on viscosity remains constant, as frequency is increased, until a certain critical frequency is reached, and thereafter decreases with frequency, until it becomes too small to be detected. The critical frequency decreases as the distance between the electrodes increases and increases as the temperature is raised.

There is no effect on viscosity with a field parallel to the lines of flow.

All these experimental facts can be explained on the theory that the effect is due to the accumulation of ions that takes place in the neighbourhood of the electrodes when the liquid conducts. The ions act as centres round which polar molecules cluster. Thus for an effect to appear it is necessary to have a liquid which conducts and which is made up of polar molecules. The limiting increase of apparent viscosity which takes place at high fields is due to the fact that the increase of viscosity of the liquid near the electrodes slows down the ions, and thus tends to diminish the number of excess ions. The variation of the effect of an alternating field with frequency also finds a ready explanation in this theory. The theory has been made to account for various other observed facts.

PART I. EXPERIMENTAL METHODS AND RESULTS

INTRODUCTION

It is now generally accepted that the variation of viscosity of normal non-associated liquids with temperature is closely expressed by the formula*

$$\eta = Ae^{b/T},$$

where kb is a factor which measures the particular energy involved in the interchange which gives rise to the viscous forces, k being Boltzmann's constant. This energy may depend upon the volume: Andrade has attempted to allow for this by taking b as inversely proportional to the atomic volume, deriving the formula

$$\eta v^{\frac{1}{3}} = Ae^{c/vT},$$

which further takes account of the effect of the thermal expansion on molecular distances. Andrade (1934, 1940) has suggested that the activation energy in question can be represented as an energy of orientation in the local electric field. Later, Eyring (1936), applying the theory of reaction rates to the viscosity problem, took b as being the activation energy necessary if a molecule is to pass over a barrier from one potential minimum to another. Whatever view be taken, it is clearly of interest to see if this energy factor can be influenced by an external electric field.

We may write

$$\eta = Ae^{E/kT}$$

or

$$\frac{d\eta}{\eta} = \frac{E}{kT}\frac{dE}{E}.$$

The value of E/kT is for most organic liquids somewhere about 3, E/k being about 800, or $E = 10^{-13}$ ergs approximately. Taking the external field as 30 kV/cm. = 100 e.s.u./cm., and considering a dipole moment of the order 10^{-18} e.s.u., then the change of energy is of the order 10^{-16} erg and $d\eta/\eta = 3 \times 10^{-3}$ which is easily measurable. By choosing a liquid of large dipole moment it should be possible, if fields of 100 e.s.u./cm. and more are practicable, to get something approaching a 1 % effect.

Turning to a non-polar molecule, the polarizability is given by

$$\frac{4\pi}{3}N\gamma = \frac{n^2-1}{n^2+2}\frac{M}{d},$$

where N is Avogadro's number, M is the molecular weight, d the density, n the refractive index and γ the polarizability. The value of $\frac{n^2-1}{n^2+2}\frac{1}{d}$ is round about 0·3. Taking again the field to be 100 e.s.u./cm.

$$\gamma = \frac{1}{4}\frac{100}{6 \times 10^{23}}0{\cdot}3 = 1{\cdot}25 \times 10^{-23},$$

---

* This formula was put forward independently by Andrade (*Nature*, 1 March 1930) and S. E. Sheppard (*Nature*, 29 March 1930). For its history see E. N. da C. Andrade, *Phil. Mag.* 1934, **17**, 703.

and energy due to field f is

$$\gamma f^2 = 1{\cdot}25 \times 10^{-19} = dE = \text{change of energy in field.}$$

Then $$\frac{d\eta}{\eta} = \frac{E}{kT}\frac{dE}{E} = \frac{1{\cdot}25 \times 10^{-19}}{1{\cdot}37 \times 10^{-16} \times 300} = 3 \times 10^{-6}.$$

There is little hope of detecting an effect of this magnitude without great refinement of any technique at present in use, and it may be noted that such a change corresponds to a temperature variation of a few ten thousandths ° C.

Many previous workers, in papers to be discussed later, have attacked the problem of the effect of an electric field on viscosity, but, in the case of polar liquids, there have been wide divergencies in the results obtained. In the experiments to be described we have made an extended series of careful measurements on the effect, with both polar and non-polar liquids. We have used much higher fields than previous workers and, in addition to steady fields, have made systematic measurements with alternating fields of various frequencies, which have thrown light on the phenomena. In particular, we have in all cases measured the small current carried by liquids under the different steady applied potentials, which has usually been neglected. Both the transverse and the longitudinal effects have been investigated, although the measurements on the latter are comparatively few, since the results, negative in all cases, showed that it had a more restricted interest, at any rate in these preliminary investigations.

In the case of polar liquids our work leads us to believe that the large effects found by many previous workers, including changes of many tens of per cent, sometimes of over a hundred per cent, in viscosity, are spurious, being caused, in the case of liquids which conduct, by an accumulation of ions in the neighbourhood of the electrodes, which leads to a modification of the state of aggregation of the polar molecules and a consequent increased viscosity in this region. This effectively reduces the channel. The local accumulation of ions also modifies the distribution of the electric field.

Our experiments further throw light on the conduction of electricity by certain classes of polar liquids. We believe that we have established the conditions under which a true effect of electric field on viscosity can be sought with polar liquids, but circumstances which need hardly be specified* have prevented us from embarking on determinations under these conditions. We hope that an opportunity to do this will occur in the future.

EXPERIMENTAL METHOD

After certain preliminary experiments in which the liquid flowed longitudinally between concentric cylinders, separated by a small gap, it was decided to use a long gap of rectangular cross-section. The width of the cross-section of the gap was large compared to the height, and the potential difference was applied between the larger opposite faces, giving a field transverse to the direction of flow.

---

* The work here described was essentially completed in 1939. A preliminary announcement of certain results was made in *Nature* (Andrade & Dodd 1939*a*, *b*).

A serious difficulty in the way of constructing an apparatus of this kind, in which metal surfaces have to be fixed in liquid-tight contact to insulating strips, is to find a cement which is a good insulator and is not attacked by the various organic solvents with which it is desired to make measurements. Another difficulty is to make the gap, which should only be a fraction of a millimetre across, sufficiently uniform. In the final form of our apparatus we utilized, for one side of the gap, three high-precision steel gauges, made by Messrs Pitter, 14 cm. long, 1 cm. wide, and 3 cm. deep, bolted together with the 14 × 3 faces in contact so that the 14 × 1 cm. surface of the central one projected 3 mm. beyond the like surfaces to either side, all these surfaces being made accurately parallel. The projecting plate was tapered away at each end and holes were drilled in it normal to the 14 × 1 face, to provide admission for the liquid. For the other side of the gap a steel base-plate 14 × 6 × 2 cm. was used. To separate the stepped face from the flat plate two quartz slips 14 × 1·5 × 0·32 cm., optically worked by Messrs Hilger, were employed. The gap was then formed by assembling the gauges and quartz strips as shown in figure 1*a*, which represents a central transverse section, the depth of the gap being 0·20 mm. Figure 1*b* shows the general arrangement of the apparatus in longitudinal section.

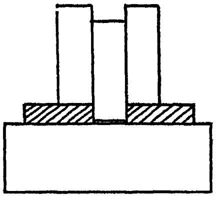


FIGURE 1*a*

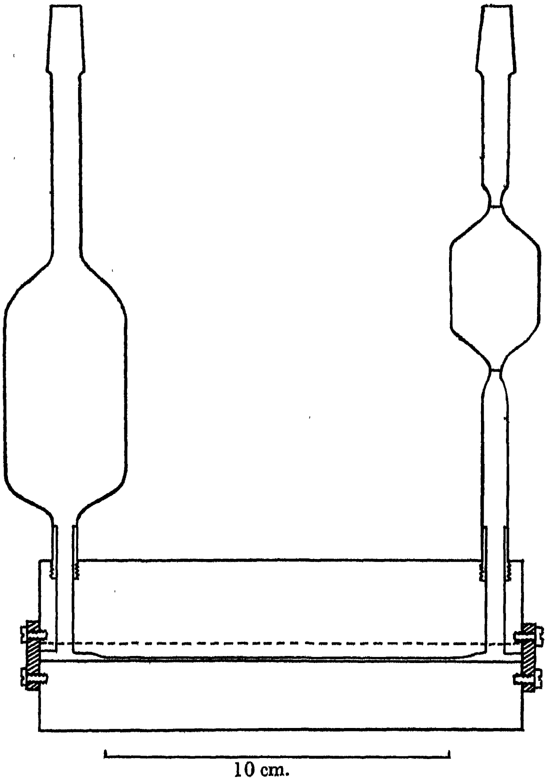


FIGURE 1*b*

The steel and quartz surfaces were so well worked that good optical contact was secured along the whole quartz-steel surface, the adhesion being such that it was extremely difficult to produce relative movement between the parts. As a precaution against displacement a cement of glue and black treacle was applied along the external re-entrant angle, but this cement played no part in retaining the liquid or in the insulation. It was clear in subsequent experiment that no liquid ever penetrated between the surfaces in optical contact. To close each end of the gap a mica plate was placed over the small rectangular hole, and held in position by a stout bakelite plate, secured by screws to the metal plates. This gave perfect insulation and the whole assembly was free from leak. A liquid left in the apparatus for several days maintained its original resistance, and other tests also showed that it was unchanged.

Stainless steel tubes were fastened with bakelite cement into the bottom of glass viscometer bulbs, of the form shown in figure 1*b*, and these tubes were screwed into the holes in the central gauge block. The viscometer bulbs were provided at their upper ends with conical ground joints, to enable them to be connected to a hand suction pump through a two-way tap. The liquid could thus be drawn to the desired height in either bulb, and flow then initiated by turning the tap so as to connect the air above the two bulbs. Two similar viscometers were constructed on these lines, so that one could be used without field while the other was used with field: the lower plates of these two viscometers were in contact with a thick metal plate, and the effect of any small, slow temperature fluctuations was thus minimized. The glass-walled tank containing the two viscometers was totally immersed in a thermostatically controlled water-bath. The variations of temperature were so small that the time of flow of a typical liquid was constant in successive runs to within 0·2 sec. for a total time of 500 sec. The lower plates of the viscometers were earthed, and the upper plate of either could be connected to a source of high potential.

The time taken for the liquid to fall between the marks on the 50 cm.^{3} viscometer bulbs was measured with a $\frac{1}{10}$ sec. stop watch and taken as a measure of the viscosity. This time was never less than 4 min. and generally much longer. Several runs were always taken and the average adopted.

The source of potential for the steady field was a number of ordinary wireless high-tension batteries. The difference of potential was measured by a Wulf bifilar electrostatic voltmeter up to 300 V, and above that by a Leybold unifilar instrument. For the alternating field a large Sullivan oscillator was used.

The current carried by the liquid, which has been neglected by most previous workers, was measured by means of a sensitive galvanometer or, when larger, by a microammeter.

For the longitudinal field, a viscometer was constructed which consisted of ten channels in series, as shown in figure 2, arranged normally to metal electrodes separated by a distance of 6 cm. The channels were 1 cm. wide and 0·03 cm. deep, cut centrally in bakelite strips which were cemented together at the edges with bakelite cement. The liquid passed up one channel and down the next, through a comparatively open space joining the adjacent ends of each pair of adjacent channels. To apply the field two large circular brass plates were cemented one to each opposite end of the channel block, so that the field ran parallel to the flow. The distance between these plates was 6 cm. Exit and entrance ports were provided at the ends of the channel system. The viscometer bulbs attached to the entry and exit tubes were of the same type as used in the transverse apparatus. The whole was immersed in transformer oil. A p.d. of 100,000 V or so was furnished by a powerful Wommersdorf influence machine.

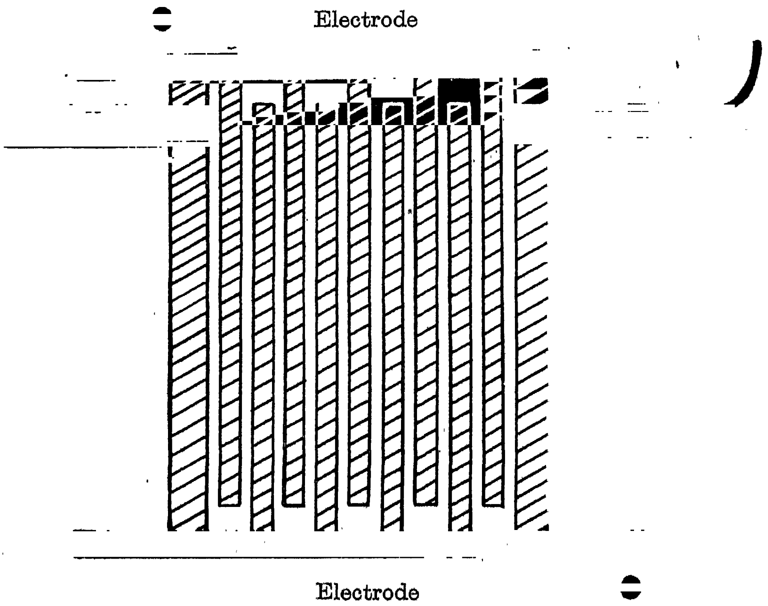


FIGURE 2

MEASUREMENT OF THE CONDUCTIVITY

In the course of the work the conductivity of certain organic liquids was often taken as an index of the purity and special measurements were therefore made in a conductivity cell. This was of glass, holding about 1 cm.^{3}, with platinum plate electrodes of area 2 cm.^{2}, 3 mm. apart. The cell was closed by a glass stopper with a mercury seal. The cell constant in the equation

$$\text{conductivity} = \frac{\text{cell constant}}{\text{resistance}}$$

was found to be 0·085 by measuring the resistance when filled with N/1000 potassium chloride solution, allowance being made for the conductivity of the water used. All measurements were made at 0° C. The resistance was measured by means of an a.c. bridge incorporating a Wagner earthing device, as described in the standard books.

PRELIMINARY EXPERIMENT. TRANSFORMER OIL

Preliminary experiments were carried out with transformer oil in an earlier version of the apparatus described, made with brass plates cemented with sealing wax to quartz separating strips. The gap was about 0·8 mm. deep. There were two exactly similar viscometers, one used as a control to eliminate the effect of small temperature variations during the run. Readings of the rate of flow were taken with both viscometers simultaneously, first with the field applied to one apparatus and later with it applied to the other. The average time of flow with field was compared with the average time of flow without field.

In the first experiments, with the oil as supplied by the manufacturer, small increases of viscosity, sometimes as great as 1 %, were noted. The oil was then degassed by maintaining it at 46° for an hour in a flask continuously evacuated by means of a rotary oil pump. Dry air was then admitted from phosphorus pentoxide drying tubes, and the oil was filtered through a fritted glass filter to remove any particles. The oil treated in this way showed no change in viscosity as great as 1 part in 2000, even when the apparatus was subjected to a potential difference of 4000 V, corresponding to a field of 50 kV/cm. The slight increases previously obtained were probably due to the presence of dissolved gas and particles.

The dipole moment of the oil was determined by finding the dielectric constant and density at four temperatures over a range of 40° C. The molar polarization was found to have the constant value of 95·5 cm.[3] Since this method is not strictly applicable to liquids, Debye's method of measuring the polarization of different concentrations of the liquid dissolved in a non-polar solvent was used. Benzene was taken as the non-polar solvent and the polarizations of four solutions of oil of different concentrations were found from dielectric constant and density measurements carried out at constant temperature. We have to thank Dr H. R. Sarna for putting at our disposal for these measurements a special condenser and bridge circuit which he was using in his own research. The molar polarization is

$$P_{12} = \frac{k-1}{k+2}\frac{M_1 f_1 + M_2 f_2}{d},$$

where M_1 and M_2 are molecular weight of benzene and oil respectively, k and d are the dielectric constant and density of the solution, and f_1 and f_2 are the molar concentrations of benzene and oil respectively. By plotting P_{12} against the concentration f_2, the values of P_2, the molar polarization of the oil, were deduced for various values of f_2. From a second graph of P_2 against f_2 the value at infinite dilution was found.

The refractive index of the oil was determined with a Zeiss-Pulfrich refractometer, and from this a value of the distortion polarization P_D of the oil found by use of the Lorentz-Lorenz equation. The dipole moment $= 0{\cdot}0127 \times 10^{-18}(P - P_D)\,T$. Our values for P and P_D were 84·9 and 85·2 c.c. respectively, from which it appeared that the oil had no measurable dipole moment. Thus for this particular non-polar oil the change of viscosity with field must be less than 1 part in 10^5 per kV/cm.

MEASUREMENTS WITH NON-POLAR LIQUIDS

Carbon disulphide

Carbon disulphide, a non-polar liquid which can be readily obtained in a fairly pure state, was the first definite chemical compound tried. The liquid was carefully distilled and then filtered through a fritted glass funnel. Both viscometers, of the construction described under 'Experimental method', were dried out, filled with the liquid and placed in the thermostat. After 4 hr. all times of flow had become constant.

About 100 V were applied to the first apparatus and the difference of the times of flow again found. The applied voltage was gradually increased by increments of about 50 V, and although voltages up to 635 V were applied, the difference in the two times of flow remained the same to within 0·2 sec.

The results obtained for carbon disulphide are given in full, as being typical of the method used for all liquids (table 1). Rough statistical treatment of the results indicates that the viscosity does not change by more than 1 part in 20,000 for an average field of 5 kV/cm., supposing the effect to be proportional to the field, while the four last determinations, made at an average field of 26 kV/cm., give a most probable variation of 1 part in 5000. It may be concluded generally that the effect of an electric field on the apparent viscosity* of carbon disulphide certainly does not exceed 1 part in 5000 for a field of 26 kV/cm., and is probably much less.

TABLE 1. TRANSVERSE ELECTRIC FIELD, CARBON DISULPHIDE

voltage on app. no. 1	galvanometer deflexion (cm.)	time of flow t_1 in app. no. 1 (sec.)	time of flow t_2 in app. no. 2 (sec.)	galvanometer deflexion (cm.)	voltage on app. no. 2	t_1-t_2 (sec.)
—	—	269·9	249·4	—	—	20·5
—	—	269·7	249·3	—	—	20·4
105	0·7	269·8	249·3	—	—	20·5
—	—	269·9	249·3	0·6	105	20·6
116	0·9	269·8	249·4	—	—	20·4
—	—	269·9	249·4	0·9	116	20·5
155	1·1	269·7	249·3	—	—	20·4
—	—	269·6	249·1	1·2	155	20·5
210	1·7	269·6	249·3	—	—	20·3
—	—	269·8	249·4	1·7	210	20·4
300	2·9	269·9	249·4	—	—	20·5
—	—	269·8	249·4	2·8	300	20·4
385	3·2	269·6	249·2	—	—	20·4
—	—	269·7	249·2	3·2	385	20·5
480	7·6	269·8	249·4	—	—	20·4
—	—	269·8	249·3	7·6	480	20·5
635	14·4	269·7	249·3	—	—	20·4
—	—	269·8	249·2	14·3	635	20·6
—	—	269·7	249·3	—	—	20·4

Gap across which voltage is applied = 0·2 mm.
Galvanometer constant = $8\cdot3 \times 10^{-9}$ amp./mm.

---

* By 'apparent viscosity' we denote the viscosity needed to give the observed time of flow on the assumption that the viscosity of the liquid is uniform throughout the gap. The reason for this term will become apparent in the discussion at the end of the paper.

Carbon tetrachloride

A pure dry specimen of carbon tetrachloride showed no change in time of flow for applied fields up to 40 kV/cm. The current passed was very small, never exceeding 0·5 μA. When the carbon tetrachloride was saturated with water there was likewise no change of time of flow with electric field, but the current was of the order of 1 mA.

Benzene

With a pure dry specimen of benzene no change of apparent viscosity was obtained even when fields of over 40 kV/cm. were applied. The current never exceeded 0·5 μA. With wet benzene, as with wet carbon tetrachloride, there was likewise no viscosity effect, but the current was once more of the order of 1 mA.

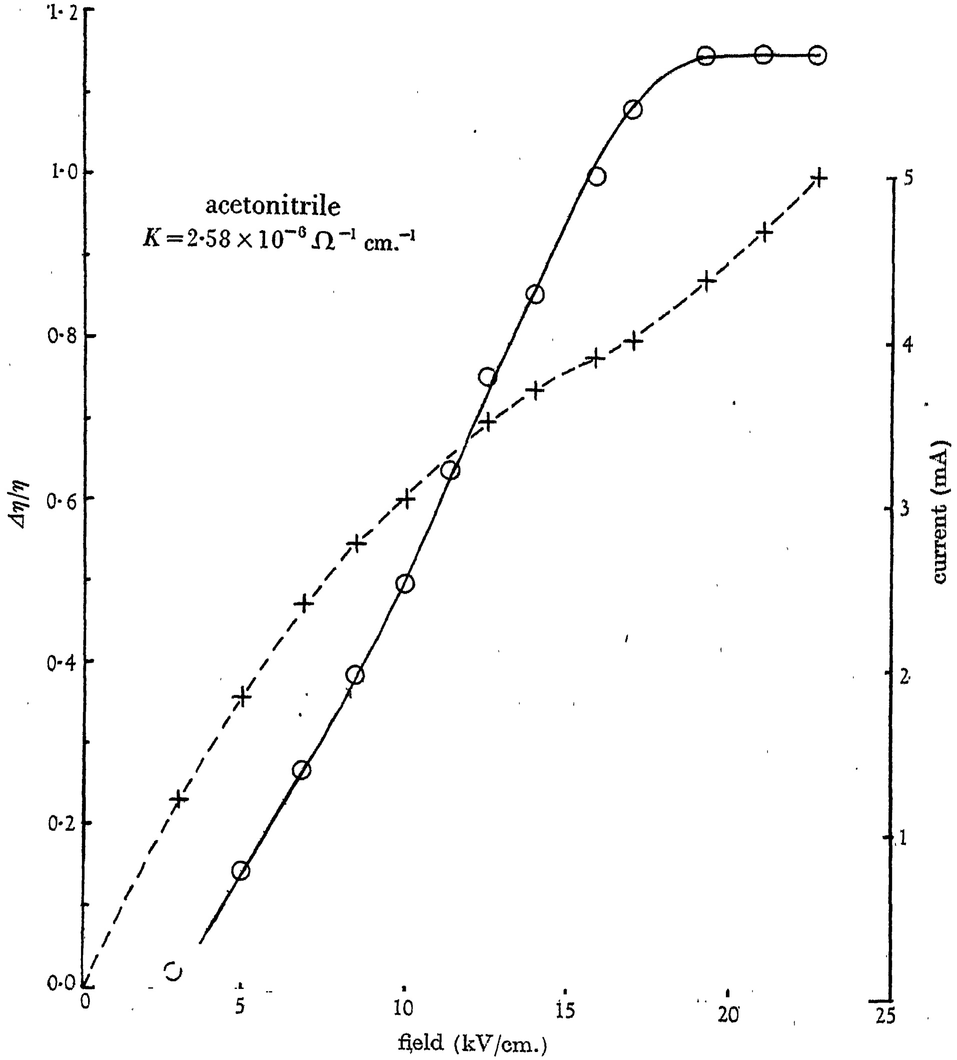


FIGURE 3a

MEASUREMENTS WITH POLAR LIQUIDS

Acetonitrile

As the first example of a polar liquid with a large permanent dipole moment, we chose acetonitrile, using the pure Kahlbaum product. The two viscometers were used, and runs made as before. An increase of time of flow with field was obtained. Interpreting this as due to rise of viscosity, the variation of viscosity with mean field* was as shown in figure 3*a*. It will be seen that very large increases were obtained, rising to a limiting value of some 115 % with the highest fields. The current for various fields is shown by the broken line, the scale being given on the right. The K given in the diagram indicates the conductivity in reciprocal ohm cm., obtained from independent measurement made as described.

The relatively large value of the current might, it was held, be due to impurity. The Kahlbaum product was therefore shaken for one hour with phosphorus pentoxide and distilled. The distillate was again shaken with pentoxide and redistilled, and this distillate again distilled four times. With this product the results recorded in figure 3*b* were obtained. The form of the curve of viscosity increase and of the current curve were similar to those obtained with the original liquid, but the effects were reduced.

Further purifications were then effected, the measurements being repeated after each process. The results are shown in figures 3 *c*, *d* and *e*,† from which it is seen that the limiting value of apparent viscosity increases was finally reduced to about 27 %, and the cell conductivity, which is recorded on each diagram, from its original value of $25{\cdot}8 \times 10^{-7}$ to $5{\cdot}78 \times 10^{-7}$ ohm^{-1} cm.$^{-1}$.

The limiting values of the increase of apparent viscosity, as indicated by times of flow, taken from the curves of figure 3, were then plotted against the conductivity, as shown in figure 4, the figures being

$K \times 10^7$	25·8	22·3	18·4	15·1	5·78
$\Delta t/t$	1·15	0·92	0·78	0·64	0·27

Δt here denotes the excess of the time of flow t above the normal field-free value for the high fields at which it reaches a limiting value. From figure 4 it appears that this excess is proportional to the conductivity and, within the accuracy that the measurements permit, that there would be very small or no increase of apparent viscosity in a liquid so pure that it did not conduct, if such a specimen could be prepared.

Variation of the current with the field. A curve of current against field for acetonitrile shows that for this very polar liquid ($\mu = 3{\cdot}4 \times 10^{-18}$ e.s.u.) Ohm's law breaks down at fields in the region 2500 V/cm., and that the current after tending towards a

---

* By 'mean field' we denote the applied potential difference divided by the distance between the electrodes. It will appear later that the potential fall is not uniform in a conducting liquid.

† The curves of figure 3 do not lie as regularly as might be wished, those of *b* and *c* crossing other curves. We intended to repeat the measurements, but were prevented by the war. These were the earliest of the systematic measurements on polar liquids: later results, such as those on acetone, are perfectly regular.

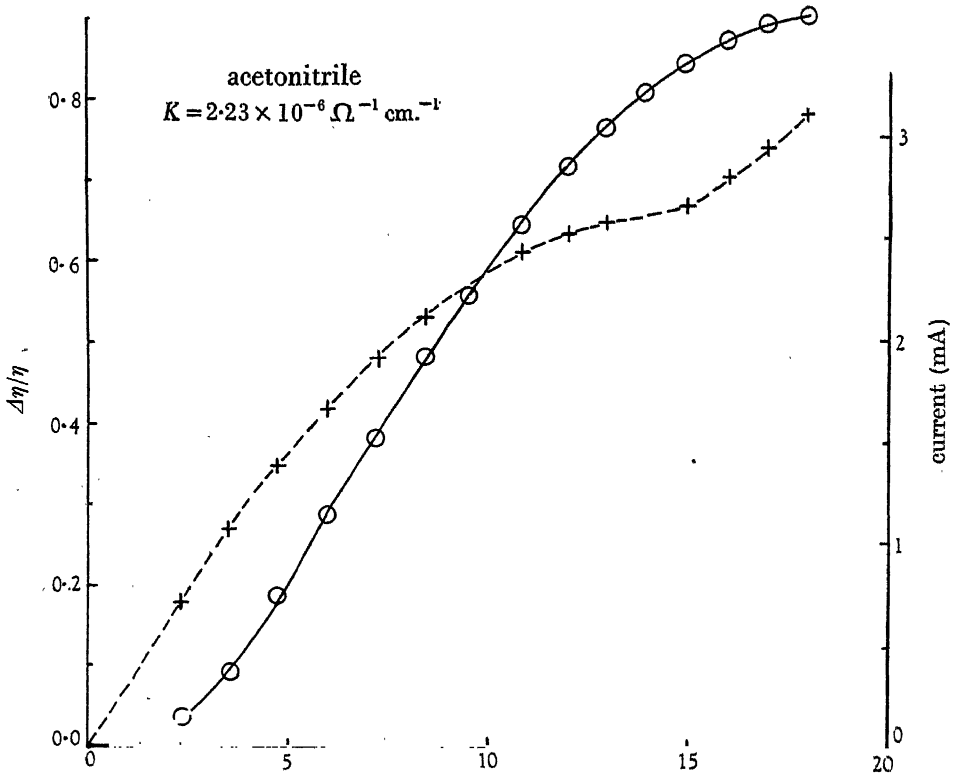


Figure 3b

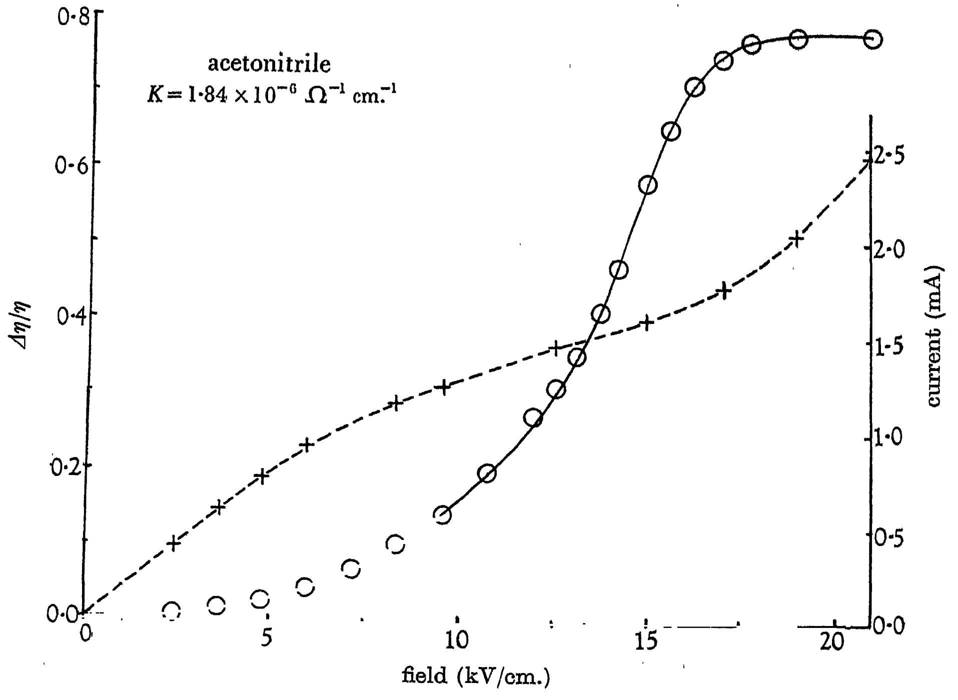


Figure 3c

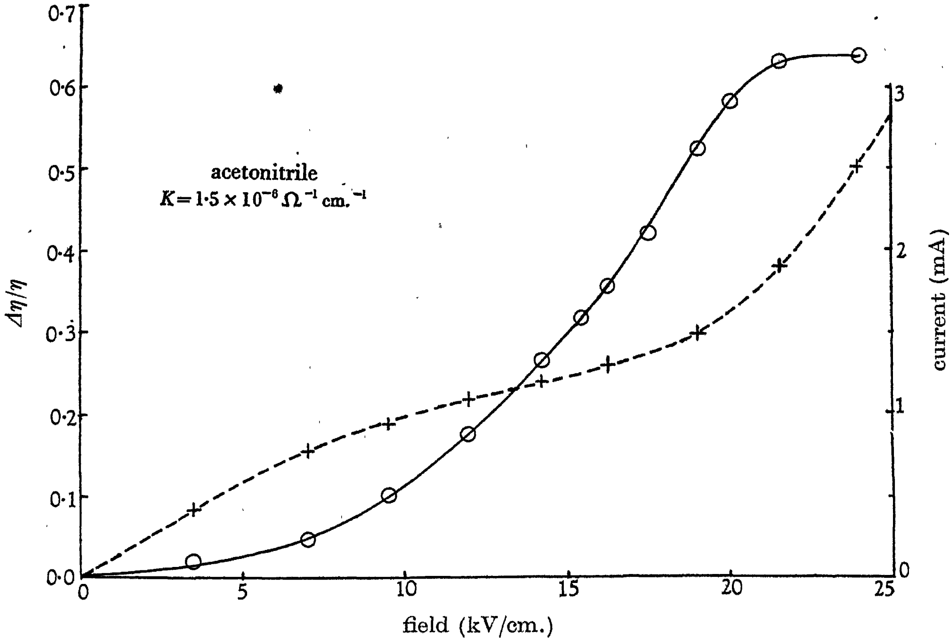


FIGURE 3*d*

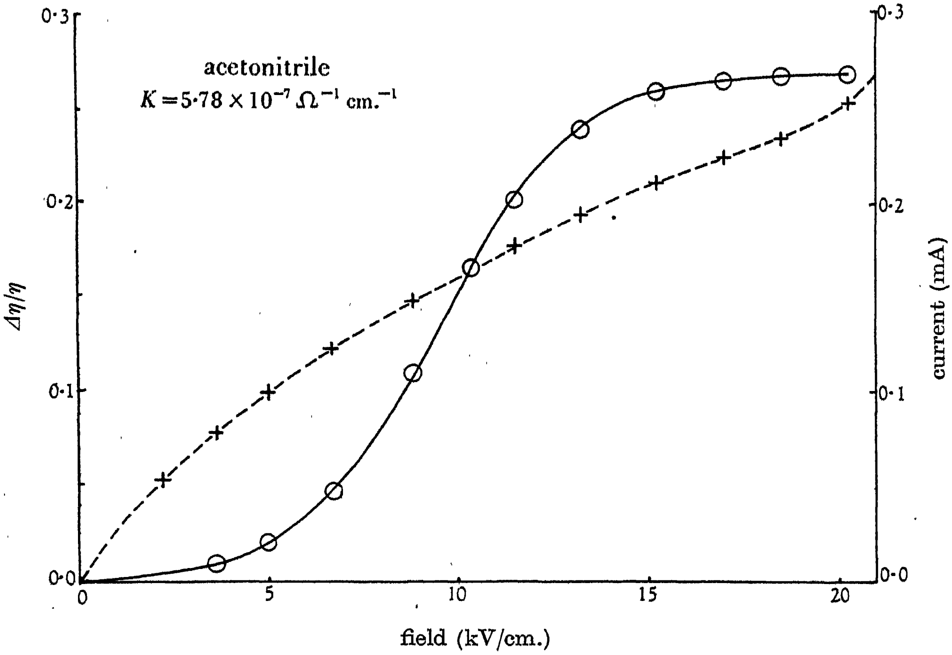


FIGURE 3*e*

saturation effect begins to increase rapidly and, of course, would become very large when a spark passed if the voltage were still further increased. Indeed, the whole curve is reminiscent of that for an ionized gas.

A very tentative estimate of the coefficient of recombination of the ions in acetonitrile can be made from these curves showing the relation between current and applied field.

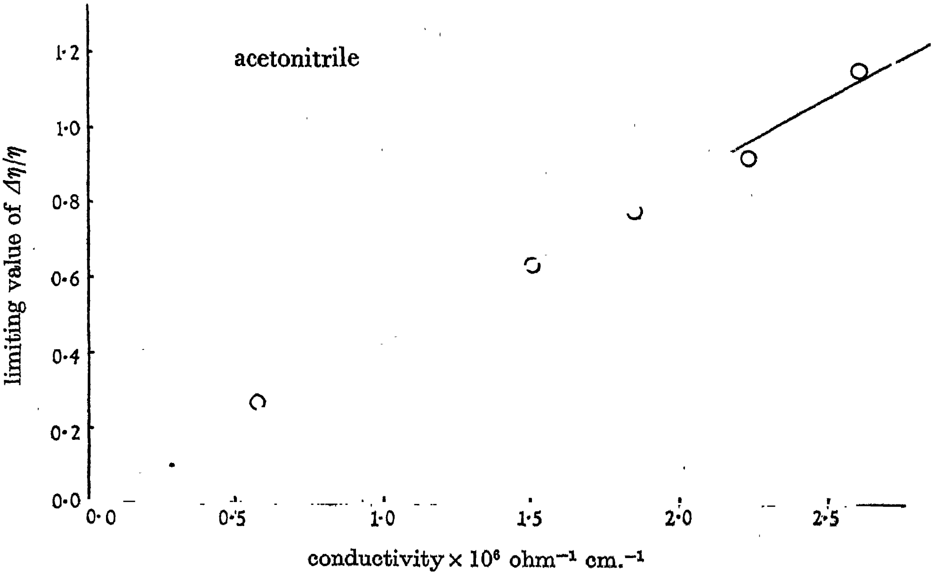


FIGURE 4

It may be shown that for an ionized gas

$$\frac{\alpha}{e}=\frac{I}{i^2}\frac{4k^2X^2}{l}, \quad \text{provided } i \ll I,$$

where α is the coefficient of recombination of the ions whose charge is e,
I is the saturation current density,
l is the distance between the electrodes,
i is the current density with an applied field X,
k is the mobility of positive and negative ions, assumed the same.

From figure 3*e*

$$\begin{aligned} I &= 0\cdot1\,\mu\text{A/cm.}^2 = 3\times10^5\,\text{e.s.u./cm.}^2,\\ i &= 1\cdot2\times10^4\,\text{e.s.u./cm.}^2,\\ X &= 250\,\text{V/cm.},\\ k &= 2\times10^{-3}\,\text{cm.sec.}^{-1}/\text{Vcm.}^{-1},\\ l &= 2\times10^{-2}\,\text{cm.}\end{aligned}$$

Therefore
$$\frac{\alpha}{e}=\frac{3\times10^5}{1\cdot44\times10^8}\,\frac{4\times4\times10^{-6}\times2\cdot5^2\times10^4}{2\times10^{-2}}=0\cdot105.$$

A value for α/e for the ions produced in liquids by rays from radium has been determined by Jaffé (1910) and van der Bijl (1912). For liquid hexane van der Bijl finds $\alpha/e = 2{\cdot}26$, which is not unreasonably different from that for the ions in acetonitrile calculated by the above method, when all the different circumstances are considered.

Heating due to the current. There must be a small heating effect due to the current, and this will in turn influence the viscosity. An estimate was made of the size of this effect. With 100 V (5 kV/cm.) the current with acetonitrile in the apparatus did not exceed 1 mA, corresponding to a generation of heat of 0·025 cal./sec. If all this heat were used to raise the temperature of the liquid, none being communicated to the metal—an extreme supposition—then for 30 cm.^{3} of a liquid of specific heat 0·5 flowing through in 300 sec., the rise of temperature would be 0·5° C. This is very much an upper limit.

That the rise was much less was indicated by the fact that the time of flow before and after the application of the field was unchanged. To investigate the matter more closely a thermojunction, protected by a very thin quartz tube, was placed both at the entry to and exit from the gap. With the field applied no temperature difference amounting to 0·01° C could be detected. A further test was made by holding the liquid steady and applying the field for 10 min. On allowing flow to start a rise of temperature at the exit of 0·3° C was observed, which rapidly fell to 0·02° C and then to zero. The first rise was due to the liquid which had been warmed in the gap for 10 min., and had communicated heat to the metal in contact with it. The entry of the cold liquid from the viscometer bulb soon cooled the metal and reduced the effect to an unobservable value.

Thus, when the liquid is flowing, even with a current of 1 mA passing, the heat exchange with the metal is so good that the rise in temperature of the liquid is less than 0·01° C. For the majority of the liquids used the current was only a fraction of a milliampere, even with 800 V between the plates. The rise of temperature is, therefore, negligible.

Acetone

With acetone, curves showing the variation of the time of flow with applied voltage were obtained for three samples of varying purity obtained by repeated distillations. These curves showed very pronounced limiting values, the increase of viscosity effect increasing by only 1 % when the voltage was increased from 240 to 360 V (see figure 5).

The conductivity of each sample was obtained by measuring the resistance in the same conductivity cell as was used for acetonitrile:

$K \times 10^6$	3·42	1·96	0·87
$\Delta t/t$ (limiting)	29·8 %	17·0 %	7·7 %

The limiting values of viscosity increase were much smaller than those obtained for acetonitrile. The graph of limiting increases of viscosity against conductivity

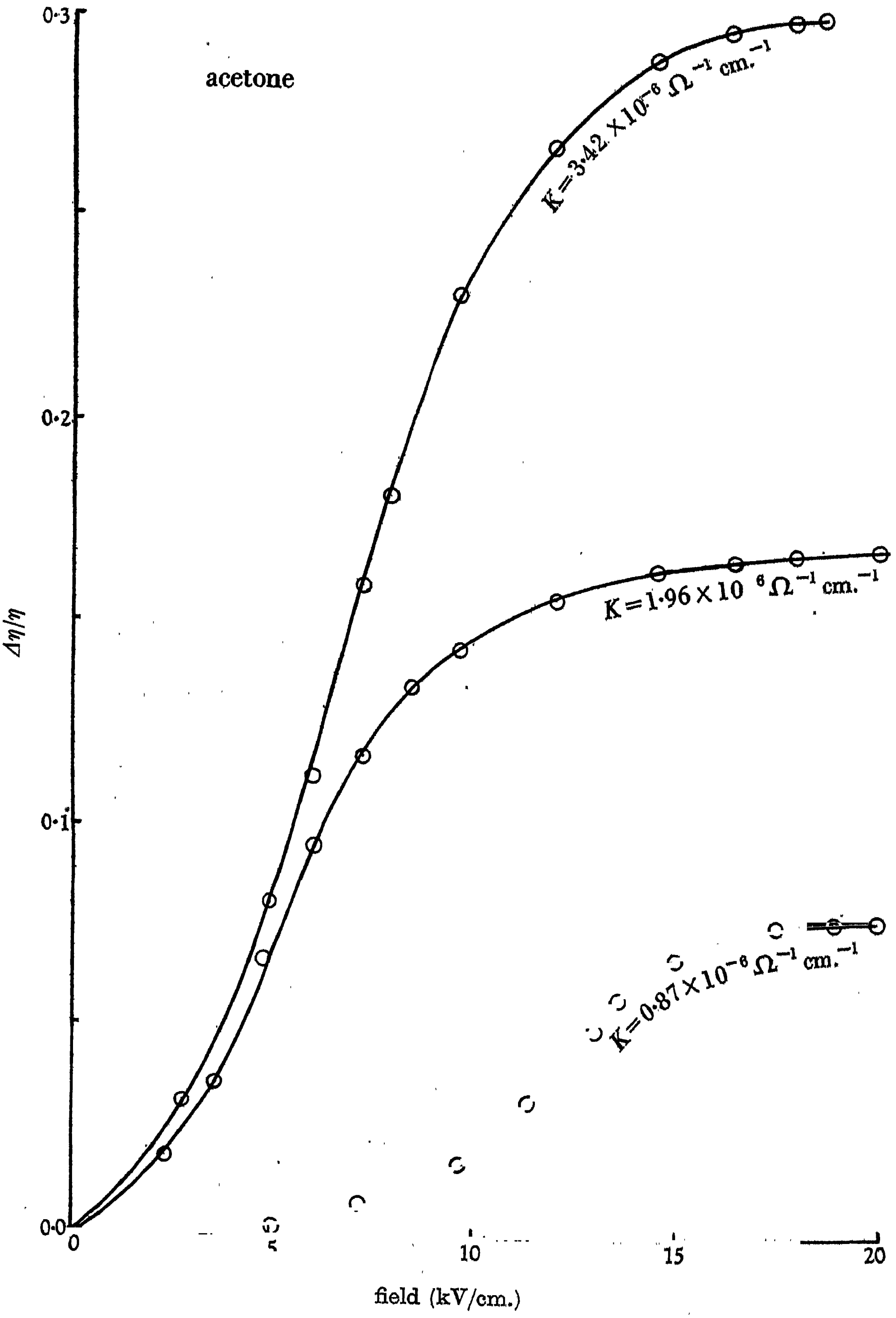

acetone
$K=3{\cdot}42\times10^{-6}\,\Omega^{-1}\,\text{cm.}^{-1}$
$K=1{\cdot}96\times10^{-6}\,\Omega^{-1}\,\text{cm.}^{-1}$
$K=0{\cdot}87\times10^{-6}\,\Omega^{-1}\,\text{cm.}^{-1}$
0·3
0·2
0·1
0·0
$\Delta\eta/\eta$
0
5
10
15
20
field (kV/cm.)

FIGURE 5

was again a straight line passing through the origin (see figure 6), indicating that there would be very small or no change of viscosity with field in a specimen of acetone which passed no current.

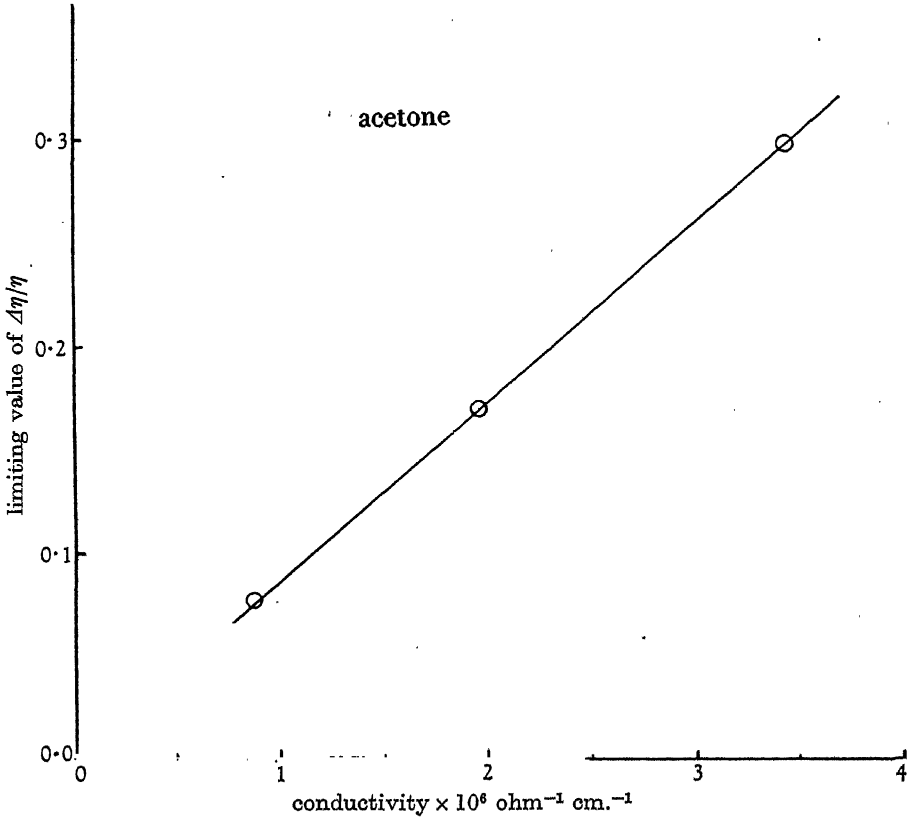


FIGURE 6

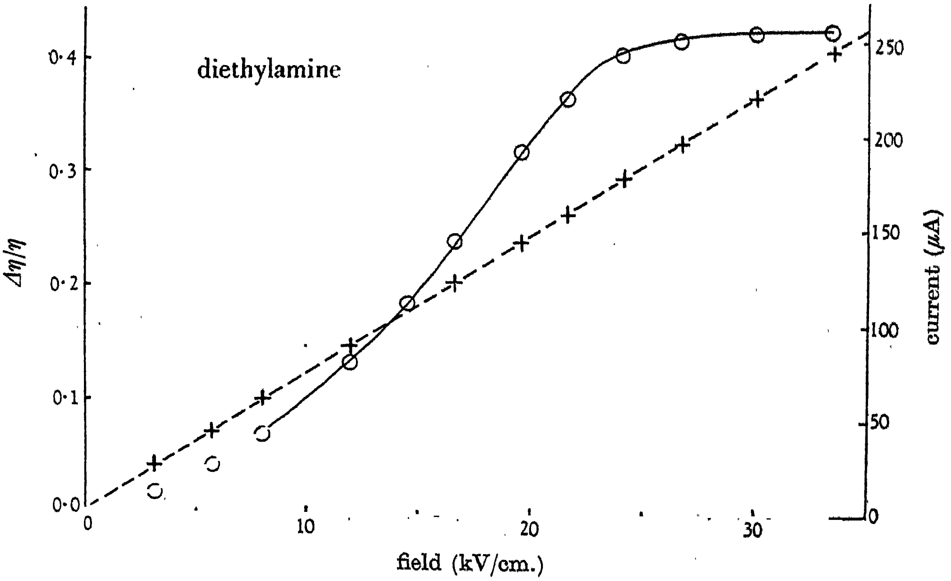
Diethylamine

A pure specimen of diethylamine was tested and was found to give a large increase in apparent viscosity. The curve representing the variation of this increase with the applied field was of the usual form, showing a well-defined limiting increase of 41·6 % as represented in figure 7. The current obeyed Ohm's law for fields up to 33 kV/cm., for which field the current was 250 μA.

Ethylene dichloride

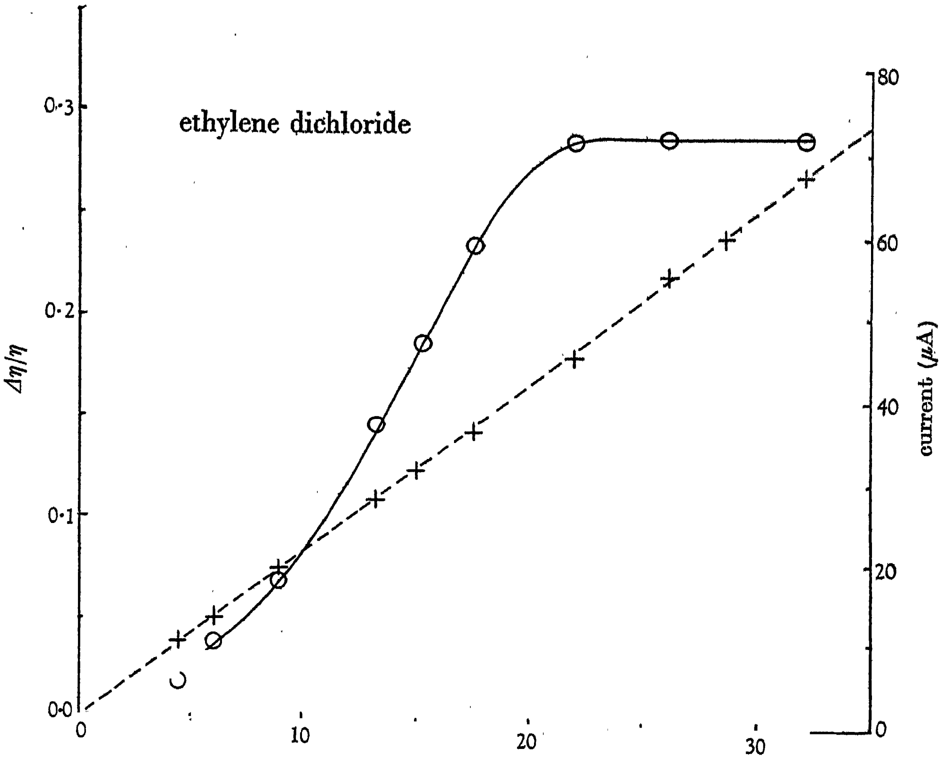
The pure dried liquid showed an increase in apparent viscosity with applied field, the curve expressing the relation between this increase and the field producing it being of the usual form. The limiting value of the increase in apparent viscosity was 28·8 % (figure 8).

The current passing through the liquid was strictly proportional to the applied voltage, having a value of about 70 μA when the field was 33 kV/cm.



field (kV/cm.)

FIGURE 7



field (kV/cm.)

FIGURE 8

The alkyl acetates

Pure specimens of the first five members of the homologous series of acetates were in turn distilled directly into the apparatus and their apparent viscosity measured under various electric fields. The curves of increase in apparent viscosity against field were of the same general shape, a limiting value for the increase being obtained for a field of about 25 kV/cm. (see figure 9). This value, however, decreased steadily

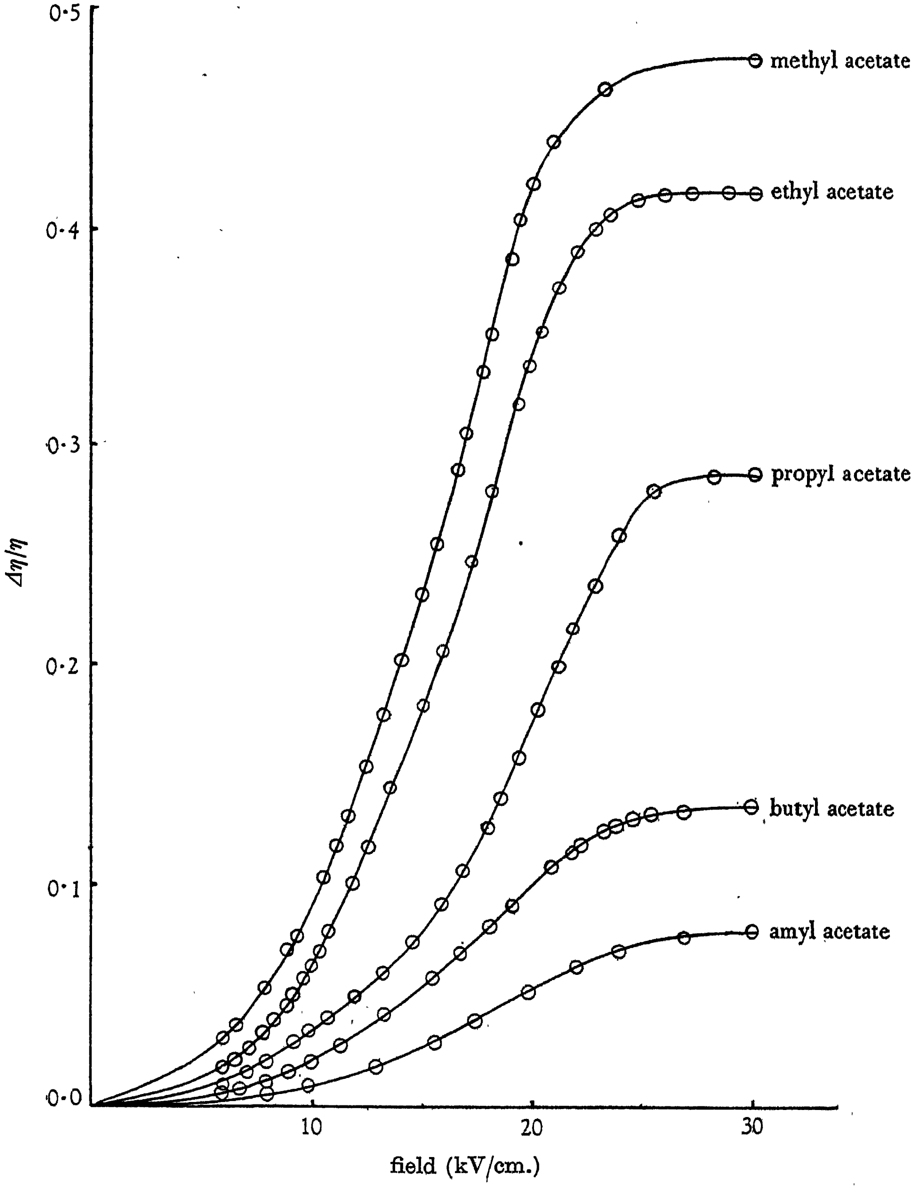


FIGURE 9

as the specimen moved higher up the series, being 47·5 % for methyl and only 7·5 % for *n*-amyl acetate. The current was much greater for the former than for the latter (table 2).

TABLE 2

substance	current for 600 V (μA)	$\Delta t/t$ (limiting) (%)
methyl acetate	1050	47·5
ethyl acetate	670	41·5
n-propyl acetate	300	29·0
n-butyl acetate	175	14·0
n-amyl acetate	85	7·5
*iso*butyl acetate	65	12·6

For the three members, normal propyl, butyl and amyl, the current obeyed Ohm's law, but for the other two members, methyl and ethyl, the current ceased to be proportional to the applied voltage for fields greater than about 10 kV/cm. This agrees with the general observations that Ohm's law breaks down when the currents are high.

Determinations were made on an isomeric form of one member of the series, with interesting results. The liquid selected was *iso*butyl acetate, and its limiting value for the increase was almost the same as that for the normal acetate, although the current was almost three times as great for the latter.

Diethyl ether

The ether used in the first place was especially pure and had been dried over sodium and redistilled. This liquid conducted hardly at all, the current never exceeding 1 μA, although voltages up to 700 V were applied. No increase in the apparent viscosity was found, even when a field of 35 kV/cm. was applied. Since an increase of only 0·2 sec. in the time of flow (400 sec.) could have been detected if present, this means that any increase in apparent viscosity must be less than one part in 2000.

The above sample was mixed with a little water, shaken up in a separating funnel and allowed to stand, with frequent shakings, for 2 weeks, by which time the ether could be assumed to be saturated with water. This fraction was poured from the funnel into the apparatus, care being taken to see that none of the excess water entered.

On applying a low field, a current of several microamperes was observed, which obeyed Ohm's law well. An increase in apparent viscosity was obtained which increased with the applied field, as shown in figure 10, but it was found that the liquid would not stand more than 9 kV/cm. With this voltage the current would become momentarily very large, resume its previous value, and then leap up again, the whole process giving indication that a spark was imminent at any moment. For fields up to 8 kV/cm., the increase in apparent viscosity was proportional to the square of the applied field, but for higher fields the effect fell off slightly, suggesting that if the field could be increased sufficiently, a limiting value for the increase would be reached, as usual. In figure 11 the fractional increase in apparent viscosity for increases up to 18 % is plotted against the square of the field. Except for the largest field applied, the current was strictly proportional to the potential difference between the two plates of the apparatus.

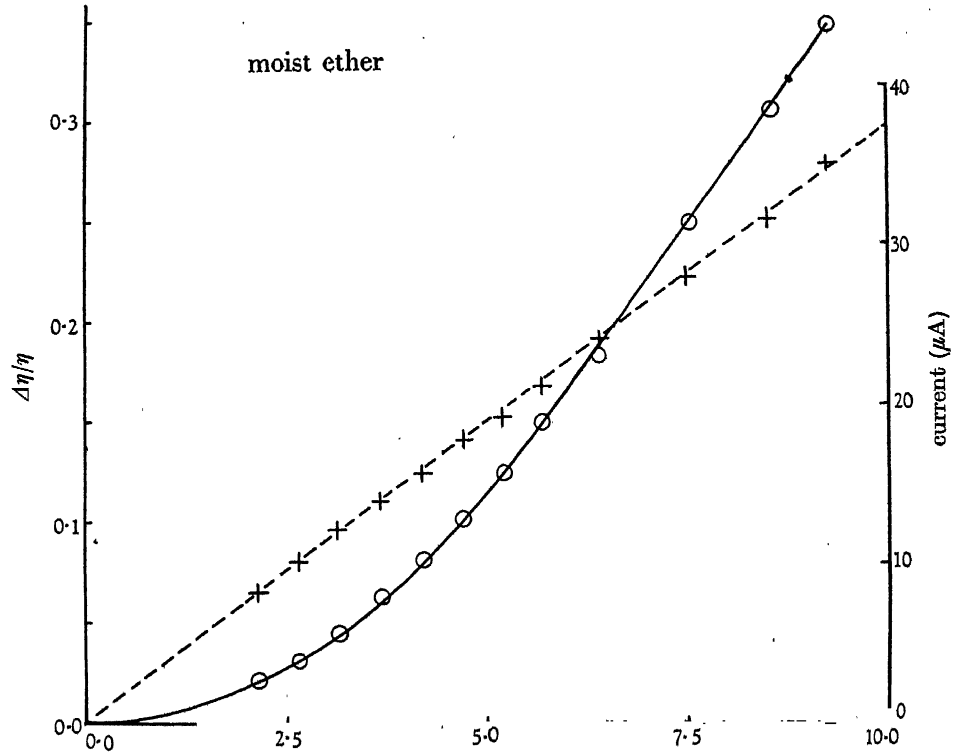


FIGURE 10

Monochlorbenzene

An exceptionally pure specimen of monochlorbenzene was obtained which passed a current of less than 0·1 μA even when the applied field was as large as 40 kV/cm. The time of flow without the field was about 600 sec., and no change in the time was observed even when the largest field was applied. Since the electrical conductivity of this specimen was almost zero, from the work on acetone and acetonitrile described earlier in this paper, no increase in viscosity would be expected even though the molecules were of a dipolar nature.

This specimen was shaken with a little water and allowed to stand in contact with it for over a month, by which time the chlorbenzene should have become saturated with water. This liquid, when placed in the apparatus, now passed a finite current when a strong electric field was applied to the apparatus. An increase in apparent viscosity of several per cent was obtained when the field was applied, a limiting effect of 8·7 % being observed under a field of 30 kV/cm., the current being about

10μA. The curve showing the variation of this increase with the applied field is given in figure 12. Up to 20 kV/cm. the increase is roughly proportional to the square of the field.

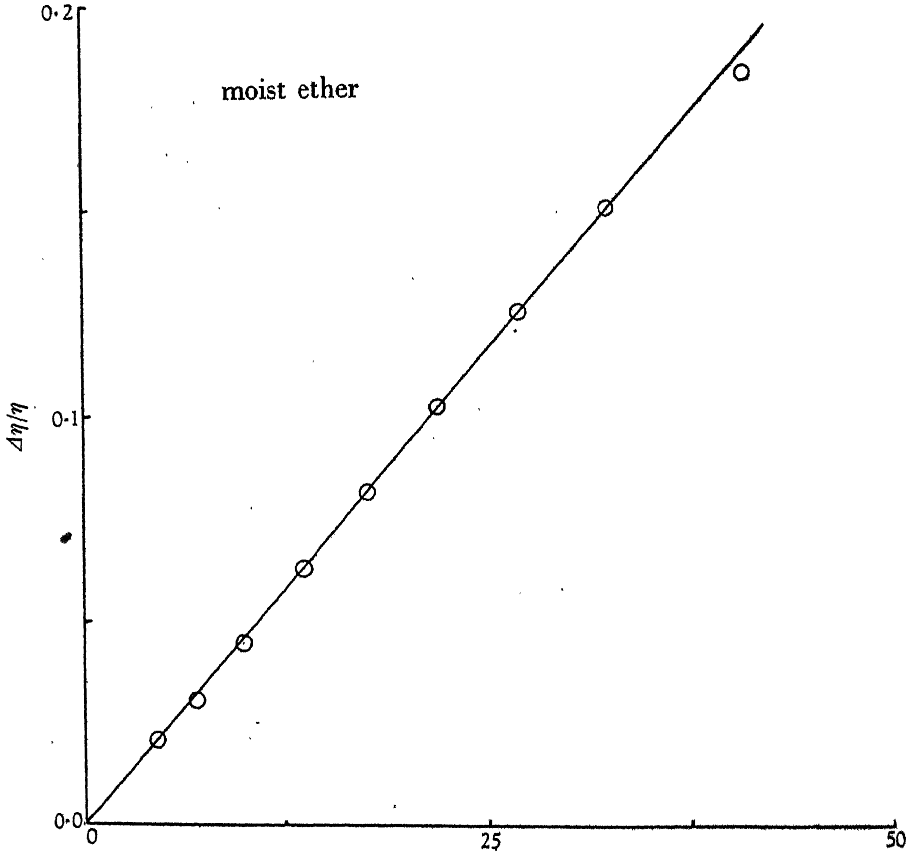


FIGURE 11

INFLUENCE OF SIZE OF CHANNEL ON LIMITING VALUES

To investigate the effect of the distance between the electrodes two fresh sets of quartz separating plates were used, giving depths of channel of 0·22 and 0·28 mm. With a pure specimen of *n*-propyl chloride the results are shown in figure 13; no measurements with this liquid were made with the 0·20 mm. gap. The limiting value of the increase of apparent viscosity was 15·3 % for the 0·22 mm. gap and 11·9 % for the 0·28 mm. gap.

With chloroform the results with the different gaps are shown in figure 14. The value of the increase of apparent viscosity at 30 kV/cm. was 31·7 % for the 0·20 mm. gap, 25·8 % for the 0·22 mm. gap and only 21·0 % for the 0·28 mm. gap. The dotted line represents the variation of the current with field, which is the same for all three gaps. It will be seen that the current is only 13μA for a field of 30 kV/cm.

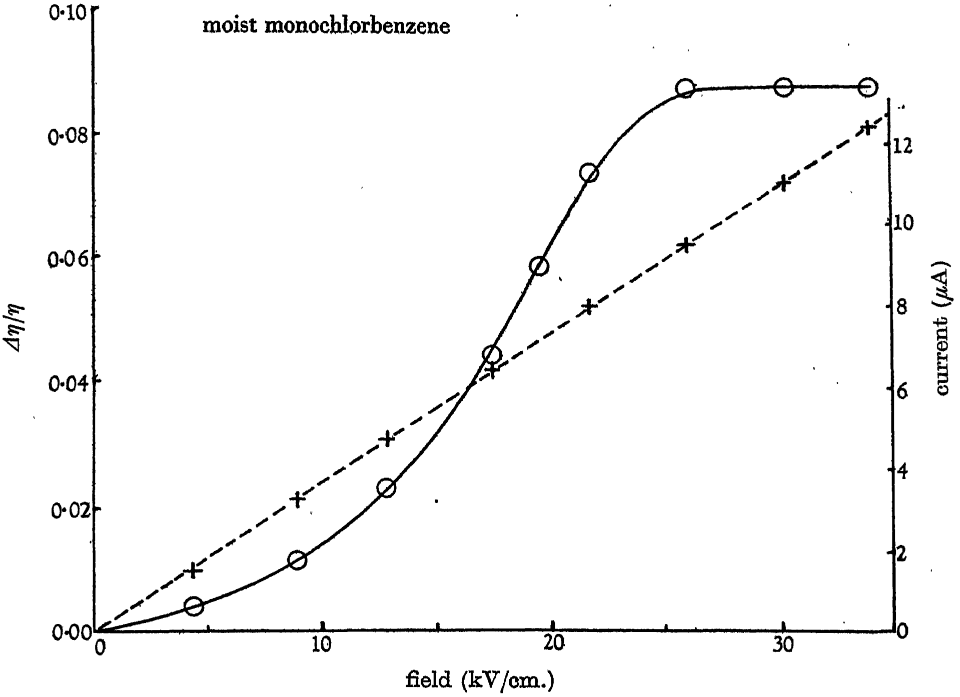


FIGURE 12

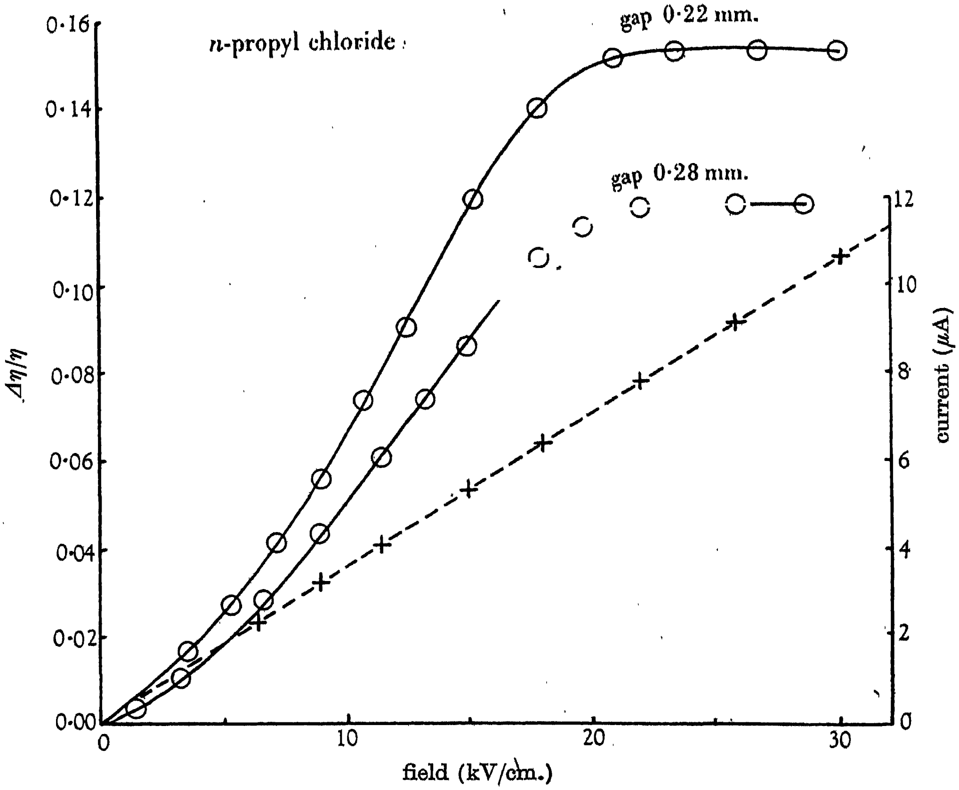


FIGURE 13

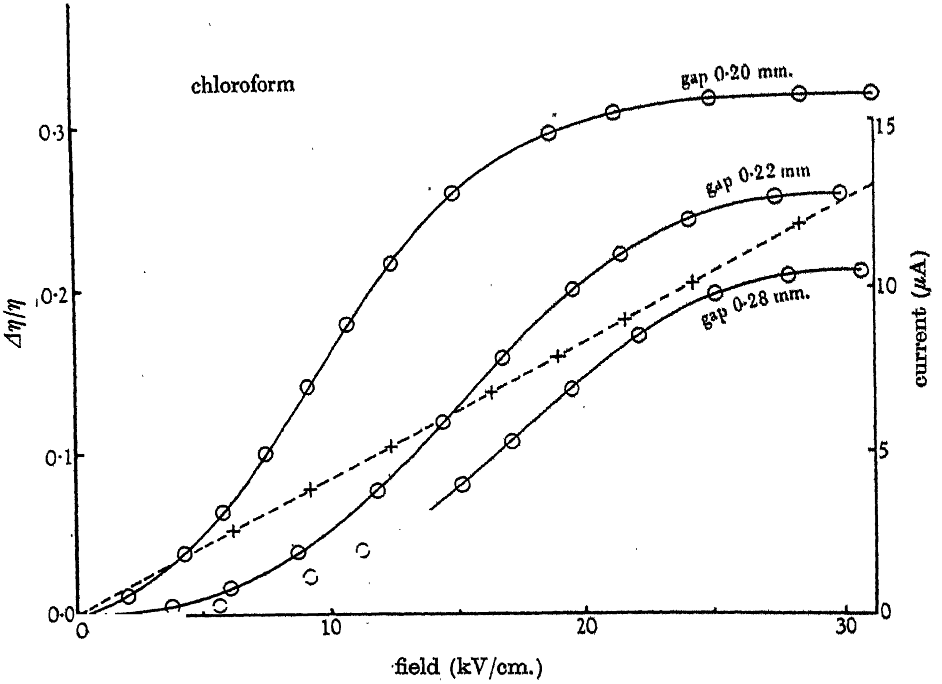


FIGURE 14

EFFECT WITH ALTERNATING FIELDS

In a first attempt to find out if any of the effects obtained could be attributed to peculiarities of the apparatus, the effect of reversing the direction of the field was tried. It was found that the increase of apparent viscosity was exactly the same for either direction, as would be expected. We then proceeded to try the effect of an alternating voltage.

In the first case the supply from the mains was used, namely, 220 V at a frequency of 50 cyc./sec., and acetonitrile was adopted as the liquid. Assuming a simple harmonic wave form and that the voltage at any instant produced the static effect appropriate to that voltage, the increase calculated was 53·6 %. The increase actually found was 52·7 %, which agrees well within the limits to be expected.

The effect of varying the frequency of the field was then investigated. The source of supply of the alternating voltage was a large Sullivan oscillator, an input of about 10 V being applied to a valve amplifier by means of a potentiometer arrangement. The output from the amplifier was applied across the apparatus, chloroform being the liquid adopted. The magnitude and wave form of the potential difference between the plates were found by means of a cathode-ray oscillograph incorporating a linear time-base circuit. The wave form was made as nearly sinusoidal as possible, to simplify the calculation of the increase in viscosity to be expected on a static basis. The peak voltage was determined by photographing the trace and then placing on the same plate three lines corresponding to zero voltage and to a known steady field applied first in one and then in the other direction.

Beginning with a low frequency, the increase in the time of flow of the liquid in the apparatus was found for a given voltage amplitude. Experiments were then carried out over a range of increasing frequencies. The increase of time of flow caused by the alternating field diminished as the frequency rose and ultimately became zero. The static curve was then obtained and the increases in apparent viscosity for the various applied alternating voltages calculated on a static basis. The actual increase was expressed as a fraction of the increase predicted on a static basis, and this ratio was found to decrease rapidly after a certain critical frequency had been reached. The results are expressed in table 3 and in figure 15.

TABLE 3

frequency cyc./sec.	$\Delta\eta/\eta$ experimental	$\Delta\eta_s/\eta_s$ predicted on static basis	$\dfrac{\Delta\eta/\eta}{\Delta\eta_s/\eta_s}$
0	0·317	0·317	1·000
50	0·210	0·208	1·009
100	0·185	0·185	1·000
200	0·241	0·243	0·990
250	0·235	0·243	0·963
300	0·149	0·243	0·615
400	0·088	0·248	0·355
500	0·045	0·235	0·191
600	0·015	0·182	0·080
700	0·015	0·195	0·077
800	0·016	0·222	0·072
1000	0·013	0·190	0·068
2000	0·007	0·192	0·037
3000	0·005	0·195	0·018
4000	0·000	0·182	0·000

gap = 0·20 mm. $T = 15°$ C

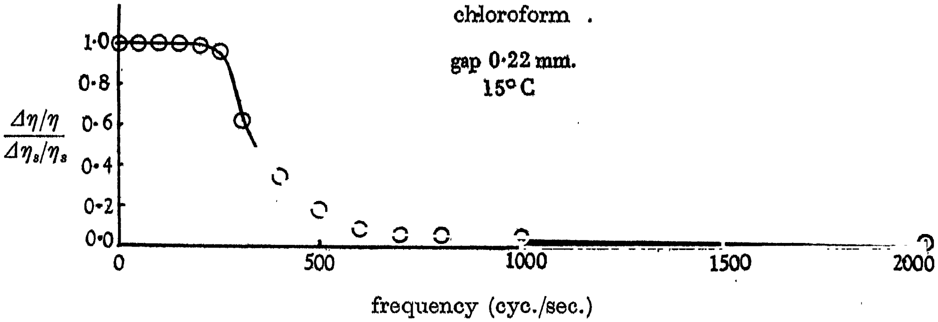


FIGURE 15

INFLUENCE OF TEMPERATURE ON THE CRITICAL FREQUENCY

Since it had been found that the increase in apparent viscosity in an alternating field diminished rapidly once a definite critical frequency was exceeded, it was decided to determine how this critical frequency depended on the temperature of the liquid.

The previous experiment was therefore conducted at 25° C, and it was again found that the full increase was obtained up to a critical frequency, but for higher frequencies the increase rapidly diminished. The critical frequency was much higher in this case, being about 450 cyc./sec. as compared with 250 cyc./sec. at 15° C (see figure 16).

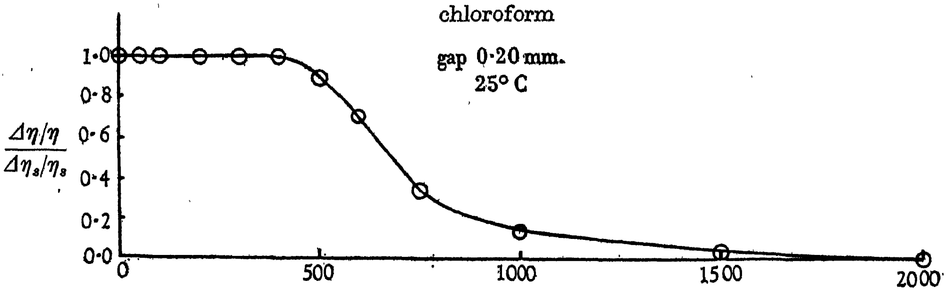


FIGURE 16

VARIATION OF CRITICAL FREQUENCY WITH SIZE OF GAP

The increase of viscosity for chloroform for various frequencies was determined in the apparatus having a deeper gap, namely, 0·28 mm.

Again it was found that the full effect was only obtained for frequencies below a certain critical frequency. The temperature of the liquid was made the same as that in the previous experiment (25° C); the value for the critical frequency was lower, being only about 350 cyc./sec., as compared with 450 cyc./sec. for the apparatus with a gap of 0·20 mm. (figure 17).

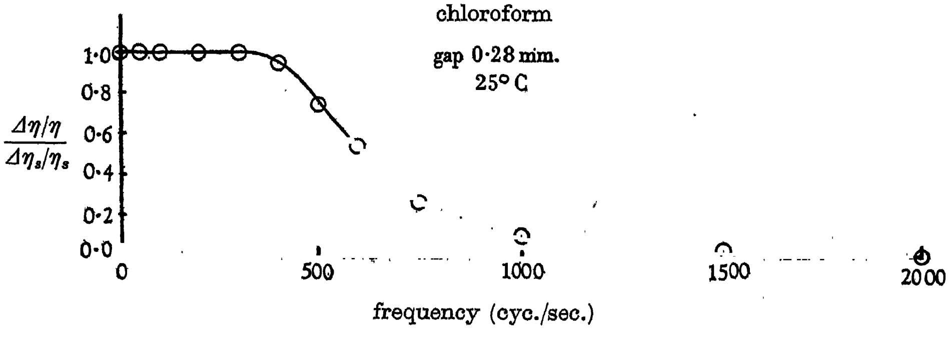


FIGURE 17

It was observed that the critical frequency was roughly inversely proportional to the size of the gap.

POLAR LIQUIDS SHOWING NO EFFECT

(1) *Toluene.* There was no change in the apparent viscosity of toluene with static field even when a field of 35 kV/cm. was applied to the apparatus. The current was small, being less than 0·5 μA with the largest field.

(2) *Metaxylene.* There was no change in the apparent viscosity of metaxylene even when fields up to 40 kV/cm. were applied to the apparatus. Again the current was small, being less than 0·3 μA with the largest field.

(3) *Anisole* (methyl phenyl ether). Fields up to 40 kV/cm. failed to produce any change in the apparent viscosity of anisole. The current was small, being only 0·8 μA under the largest field.

LONGITUDINAL EFFECT

Experiments with the direction of field parallel to the lines of flow were carried out with the apparatus described earlier, the source of potential being a Wommersdorf influence machine capable of giving about 100,000 V on moderate load. Acetonitrile was the first liquid tried, but it was found to conduct too well to allow the necessary potential difference to be established. Chloroform was then used, as being a liquid which showed a fairly large increase of apparent viscosity with a transverse field of 14 kV/cm. while conducting comparatively slightly. With an applied potential of 80 kV, giving a field of about 14 kV/cm., no change in apparent viscosity was detected. The time of flow was 250 sec.: a decrease in time of flow varying from 2 to 3 sec., according to the applied potential, was quantitatively explained by a slight rise of temperature that was found to take place. It can be said, then, that in the longitudinal case no change of apparent viscosity amounting to 1 part in 500 takes place under the conditions specified.

PART II. DISCUSSION AND THEORY

PREVIOUS WORK

The early work on the effect of an electric field on the viscosity of liquids is not very convincing. König (1885), who seems to have been the first to work on the subject, used a glass capillary tube between condenser plates about 1 cm. apart, to which a potential difference of about 14,000 V was applied. He found no change of viscosity with carbon disulphide. Duff (1896) used the falling sphere method, his spheres being of metal, and obtained an increase of viscosity with castor oil of about 0·5 % and a possible decrease with paraffin oil with a field of 27 kV/cm. The method, which entails a conducting sphere entering and leaving a strong field, is obviously open to criticism. Quincke (1897) observed the damping of the oscillations of a glass sphere in a liquid between the plates of a condenser and found, with fields not exceeding 3·3 kV/cm., a very large increase of viscosity, with, among other liquids, carbon disulphide, ether and benzene; the results are certainly spurious. Pochettino (1903) enclosed the liquid between two metal disks held separate at a distance of about 14 mm. by an ebonite ring; this arrangement was hung up by a bifilar suspension, the disks being horizontal, and the damping of the oscillations due to the enclosed liquid was measured. With about 15,000 V applied to the plates he obtained an increase of viscosity of 1·5 % with benzene and 4 % with xylol. There is, however, a total force of some 20 g. wt. between the plates with the field in question, which, in the absence of precautions, may well have produced sufficient deformation to account for the result. Ercolini (1903) used an apparatus in which the liquid was contained between insulated concentric cylinders. The whole system was suspended and the damping of the oscillations taken with and without the field of about 10 kV/cm. between the cylinders: he found no change in the viscosity of benzene within 1 %, about the limit of his accuracy. Pacher & Finazzi (1900) measured the time of axial flow between coaxial glass tubes. The potential difference was applied between mercury contained in the inner tube and tin foil wrapped round the outer one, so that the field in the liquid was uncertain but large. No effect was obtained with ether, benzene, carbon disulphide and other liquids. The early methods are all open to criticism and, in view of later results, little reliance can be placed on the measurements, a possible exception being that of Pacher & Finazzi.

The subject was not taken up again until 1933. Herzog, Kudar & Paersch (1933, 1934) used first of all an ordinary capillary tube between condenser plates, like König, but later a tube of square cross-section, two opposite walls being of glass and the other two of platinum, between which a field was established. This form is very unfavourable for qualitative work, since the viscous flow is as much influenced by the two opposite sides which are at the same potential as by the sides between which a large, but not uniform, difference of potential is maintained. We shall show that all the positive effects observed are due to a modification of the properties of the liquid in the neighbourhood of the electrodes: the square channel is unsuited to reveal such an effect. Herzog and his collaborators used a low field only, viz. 2200 V/cm., and found no effect with the non-polar liquids hexane, *cyclo*hexane and carbon tetrachloride. With ether they obtained a large increase, about 18 %, which in view of our results and those of Dobinski (1935) can only mean that the ether was very wet. With a range of polar liquids they obtained an increase of viscosity varying from something less than 1 up to 17 %. Determinations were only taken at one field strength and the results do not take us very far. Sokolov & Sosinski (1935, 1936, 1939), like Herzog *et al.*, used a tube of square cross-section, the side length being 0·6 mm., and applied fields up to 17 kV/cm. They obtained no effect with hexane, benzene, toluene and carbon tetrachloride. Having come to the conclusion that the increase in time of flow was a function of the conductivity of certain liquids, they artificially increased the conductivity by the use of HCl vapour. The vapour was passed through a given quantity of the liquid, and small quantities of this contaminated liquid were then added to the pure liquid; in this way a range of conductivities was obtained. With chloroform and ether, which were particularly investigated, the increase of time of flow was plotted against $\surd(\lambda/\lambda_0)$, where λ is the conductivity of the contaminated liquid and λ_0 that of the pure liquid. The authors also carried out experiments on the increase of time of flow (apparent viscosity) with applied potential at frequencies of 0, 50 and 1000 sec.$^{-1}$. The curves for the three frequencies cross one another in an unsystematic way for different liquids, so that nothing can be deduced from their results as to variation of the effect with frequency. They also increased the conductivity of hexane with HCl vapour, and found an effect on viscosity, which is in contradiction to our experiments on wet non-polar liquids. Their final conclusion is that the dependence of the effect on the purity of the liquid is an exceedingly complicated problem. Considering the results, they decide correctly that the effect cannot be explained by an orientation of the molecules, a direct transport of momentum by ions, or various other mechanisms which they consider. They finally come to the conclusion that the electric current causes a general convection of the fluid parallel to the electrodes, that the effect is 'a pure hydrodynamic phenomenon due to the convection of a non-homogeneous liquid placed in an electric field', although they admit that there is so far no satisfactory explanation of such movement. We consider that our results, which establish that with many liquids the current increases linearly with the field while the apparent viscosity reaches a limiting value, show that this explanation cannot be correct. Later we will bring other arguments against it. The chief significance of the work of Sokolov & Sosinski is that they draw attention to the fact that a current accompanies the effect.

Alcock (1936) used a rotating cylinder apparatus unsuited to accurate work: he took no steps to purify his liquids or to control the temperature. He applied a steady field only, not exceeding 2500 V/cm., and found no effect with carbon tetrachloride and benzene. As regards other liquids, in some cases he found first of all a small decrease and then an increase of viscosity with increasing field.* With polar liquids such as diethyl ether (presumably wet), chloroform and acetone he obtained results which he interpreted as increases of viscosity exceeding 100 % at fields of 1600 V/cm. or so.

Dobinski (1935) used the method of flow, parallel to the axis, in the space between a cylindrical metal rod 1 mm. in diameter and a coaxial metal cylinder of 2 mm. internal diameter, the potential difference being applied across the gap, which involves a non-uniform field. His field did not exceed 4400 V/cm. He found no effect with benzene, carbon tetrachloride and carbon disulphide. With dry ether there was no change of viscosity, but with wet ether (1·2 % water) there was an increase of 12 % at 2200 V/cm. Acetone likewise showed no change when dry, but an increase when wet. He found an increase of viscosity at 2200 V/cm. with chloroform, methylene chloride and nitromethane, which are polar, while, as regards aromatic polar fluids, he found no change with xylene, toluene and nitrobenzene and a decrease of viscosity with aniline and *o*-nitrotoluene. Sokolov & Sosinski have shown that these decreases are to be attributed to heating effects. Within their limited range Dobinski's results, which are mostly carried out at one field only, are in agreement with ours except for this decrease. Kimura (1937) has published a note on the effect of an electric field on a solution of stearic acid in benzene, with which he went up to 30 kV/cm.* With increasing field he obtained an increase of viscosity which reached a limiting value. We have not dealt with solutions, and merely cite this note since it is the first record of a limiting value.

---

* We suggest that the decrease was probably a temperature effect, due to heating of the liquid by the current.

Björnståhl & Snellman (1935, 1937) used a form of apparatus different from those of all other workers, viz. a suspended disk parallel to a uniformly rotating disk, and made extensive use of alternating fields. The method seems open to criticism on the ground that small asymmetries may produce spurious forces when the strong electric field is applied. In any case the results are hard to reconcile among themselves: for instance, with monochlorbenzene the authors find with a steady field of 10 kV/cm. an increase of viscosity of 189 %, while with the same field at 25 cyc./sec. they find an increase of 440 %. There seems every reason to suppose that with such slow fluctuations of field the effect must be the same as in the static case, that is, less than the static effect if the recorded voltage is the peak voltage, instead of being considerably more than twice as great. Again, with a solution of monochlorbenzene in benzene they found no effect at all, while with nitrobenzene, which gave a smaller effect in the pure state, they obtained a large increase with the solution in benzene. Further, the way in which, with nitrobenzene, the viscosity first decreased and then increased to the normal value as the voltage was raised, and the whole variation of the effect with frequency in the case of chlorbenzene and of bromobenzene, which shows, in both cases, the most capricious changes, point to some fundamental irregularity in the method of measurement. The curves do not form a self-consistent system.

The result of all this previous work is that there is general agreement among the more careful workers that there is no effect with certain non-polar liquids. Great discrepancies are shown by the results on polar liquids. There is, however, general agreement that with some polar liquids there is a very large apparent increase of viscosity with field, although complete disagreement as to the magnitude of the effect, and as to its variation with field strength. Such measurements as have been made on the effect of alternating field are unsystematic and discordant. The marked differences and inconsistencies show clearly that there are factors involved which have not been taken into account by previous experimenters.

GENERAL RESULTS OF OUR EXPERIMENTS

We have established that an electric field transverse to the flow has no effect on the viscosity of non-polar liquids. This has been found by the previous workers just cited, but we have widened the limits within which this can be stated. With transformer oil any change of viscosity must be less than 1 in 5000 for a field of

---

* Probably. The intensity of the field as given is in volts, not V/cm., and there is a contradiction between the table and the diagram as to the strength.

50 kV/cm., with carbon disulphide less than 1 in 6000 at 32 kV/cm. With carbon tetrachloride and benzene the change of viscosity is also within this limit at about 40 kV/cm.

These non-polar liquids are, when dry, very bad conductors of electricity compared to polar liquids such as acetonitrile, chloroform and others that show a large change of apparent viscosity with field. When saturated with water, however, they pass currents of the same order as these polar substances, that is, thousands of times as great as they do when dry, but still show no change of viscosity with field. This proves definitely that the viscosity effect is not a consequence of the current passing, as Sokolov & Sosinski have suggested.

A large class of polar liquids, when dry and pure, by ordinary standards, show a big increase of apparent viscosity with electric field, the increase reaching a limiting value as the field increases, and also conduct relatively well: acetonitrile (3·16), acetone (2·74), dimethylamine (0·96), ethylene dichloride (1·51), methyl acetate (1·67), ethyl acetate (1·76), propyl acetate (1·89), butyl acetate (1·84), amyl acetate (1·91), *n*-propyl chloride (1·94), chloroform (1·18). The figures in brackets give the dipole moments when multiplied by 10^{-18}. In the case of the two liquids first named it has been shown that, with specimens at different stages of purification, the limiting increase of viscosity at high fields is proportional to the conductivity, the result indicating that a specimen so pure as not to conduct at all, if it could be obtained, would show very small or no change of apparent viscosity with field.

Certain polar liquids, e.g. ether (1·15), monochlorbenzene (1·56), show no effect of an electric field on viscosity when dry, and also conduct very badly. When saturated with water they conduct much better and show the same type of increase of apparent viscosity with field as the large class of conducting polar liquids cited in the preceding paragraph. Toluene (0·4), metaxylene (0·4) and anisole (1·2) are likewise polar and neither conduct nor show a viscosity effect: the effect of wetting them was not tried.

The liquids fall, then, into three classes:

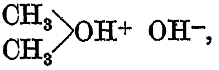
(1) Non-polar liquids, which show no viscosity effect, whether they are dry and non-conducting, or wet and conducting.

(2) Polar liquids which do not conduct and which show no viscosity effect.

(3) Polar liquids which conduct, either as ordinarily termed 'pure' or on the addition of moisture. These show the characteristic viscosity effect, approaching a limiting value at high fields. Successive purifications indicate that if the specimen could be prepared so as not to conduct at all, there would be little or no effect.

The general conclusion is, then, that for an effect to exist, neither conduction nor polar molecules alone suffices, but that both must be present.

Conduction is, of course, an index of ionization, and the conducting behaviour of our liquids is in accordance with general chemical considerations from this point of view. Toluene, metaxylene and anisole should, from the point of view of chemical structure, show no, or very little, tendency to ionization. Ether and monochlorbenzene should not ionize when dry, but may reasonably be expected to do so in the presence of water. Thus ether should ionize according to the scheme



$$\begin{matrix}CH_3\\CH_3\end{matrix}\!\!>OH^+\ OH^-,$$

while as regards monochlorbenzene, which shows a comparatively small viscosity increase, the conduction and viscosity changes may be due to the ionization of the water alone. It is to be noted that conduction and viscosity effects are both much larger with wet ether than with wet monochlorbenzene.

The liquids of class (3) ionize readily in the presence of a trace of water. Thus acetonitrile $CH_3.C\equiv N$ probably goes to $(CH_3.CNH)^+$ and OH^-. In the case of the homologous series of alkyl acetates the effect systematically decreases as we go from methyl to butyl, which is to be expected if the effect is connected with hydrolysis. Thus the connexion of the viscosity effect with ionization, shown by the current, is also in accord with general considerations of chemical structure. To know how the ionization comes about is not, however, essential for our theoretical considerations; it suffices to know that ions are there, without knowing exactly what ions they are.

Experiments with different depths of channel show that the limiting effect diminishes with increasing depth.

Fundamental for our consideration is the result that with an alternating field the viscosity effect begins to diminish when a certain critical frequency is reached, and ultimately tends to become very small or vanish as the frequency is still further increased. The critical frequency diminishes as the depth of the channel is increased, and increases as the temperature is increased.

General explanation of the effects

It has been shown by many workers that poorly conducting liquids behave under an electric field in a manner analogous to gases. Schweidler (1901 *a*, *b*, 1904), working with plates immersed in toluol, showed that the potential gradient was greatest at the electrodes, and with petroleum that the current increased more slowly than the applied potential. Jaffé (1908) and van der Bijl (1912) worked with liquids in which the ions were produced by γ-rays from radium, and found values for ionic mobilities and coefficient of recombination; van der Bijl, in particular, found that the recombination obeyed the laws valid for ions in gases. That poorly conducting liquids should behave like gases rather than like normal electrolytes is not astonishing, in view of the fact that the difference between gases and normal electrolytes is due to the relative scarcity of ions in the former.

Our results have led us independently to the conclusion that the liquids we have used conduct ionically, in the gaseous manner. Under the influence of the field, local concentrations of ions appear at the electrodes, giving rise to abnormal cathode and anode fall of potential, as widely observed in gases and as measured by Schweidler, in a particular case, with liquids. The field gradient is greatest at an electrode, and diminishes as the distance from the electrode increases until a region is reached where it is zero.

The general nature of our results can be explained by the concentration of ions in the neighbourhood of the electrodes that takes place in a conducting liquid, if one makes the supposition that such ions, in consequence of the intense local field, act as centres around which polar molecules, if present, cluster. The viscosity of a liquid containing an appreciable quantity of such clusters will exceed that of the normal liquid. On this basis it is seen, in the first place, why, for a marked viscosity effect, it is necessary to have both conduction and polar molecules, since without conduction there is no local concentration of ions and without polar molecules no clusters are formed. In the second place this general picture explains why, in the case of alternating fields of increasing frequency, after a certain frequency ν_c is reached the viscosity effect diminishes and ultimately becomes negligible. For the establishment of the full effect it is necessary for the ions to move through a distance approximately equal to the depth of the channel. When the half-period is of the order of the quantity depth-of-channel/velocity-of-ion, the effect will begin to diminish and will vanish when the period is so small that the ions do not move appreciably.

The order of the ionic mobility—nothing more can be attempted in view of the complications due to non-uniformity of field and to the fact that equilibrium is not reached—derived from this point of view turns out to be reasonable. Table 4 shows the approximate figures. These values of k are of the right order. For instance, the mobility of the NO_3^- ion in water at infinite dilution is about 6×10^{-4}, and the viscosity of water is about twice that of chloroform. For the CS_2 ion produced in CS_2 by radium radiation, van der Bijl found $k = 6{\cdot}6 \times 10^{-4}$, the viscosity of CS_2 being about 0·6 times that of chloroform. The increase of mobility with increasing temperature, indicated by the experimental results, is explained by the decrease of viscosity. Discussion of the difference of mobility found with the two different depths of channel would demand a more detailed analysis of the distribution of field under changing voltage than is possible at present.

TABLE 4

depth of channel (cm.)	peak voltage	temp. °C	critical frequency	mobility k (cm./sec.)/(V/cm.)
0·020	327	15	235	9×10^{-4}
0·020	333	25	450	17×10^{-4}
0·028	344	25	350	25×10^{-4}

The lack of a longitudinal effect is also consistent with the general theory here put forward.

Before passing on to a more detailed discussion of our theory it may be pointed out that the effects cannot be explained by the transport of momentum consequent on the transport of ions. In the first place wet CCl_4 and wet C_6H_6 carry currents of the same order as the liquids that show the viscosity effects, and have ions of the same order of mass, but show no effect. In the second place we have calculated the viscosity effect to be anticipated as due to the transfer of momentum by ions of the kind of mass in question, and with chloroform, for instance, the change of viscosity due to the drift of ions with the highest field would be about 1 part in 4000, and so quite negligible compared to the effects found experimentally.

THEORY OF THE VISCOSITY INCREASE

The greatly increased time of flow which occurs when an electric field is applied at right angles to the direction of flow we have attributed to the concentrations of free ions in the neighbourhood of the electrodes, which leads to clusters of polar molecules involving local increases of viscosity. This theory explains qualitatively the class of liquid that shows the effect and also the fact that in an alternating field the field begins to diminish at a certain frequency and vanishes at a still greater frequency. It remains to show why the increase of apparent viscosity tends to a limiting value and why the effect diminishes as the channel in the viscometer is made deeper. These problems will now be handled quantitatively, but it may be said at once that the limiting value is due to the decreased velocity of the ions in the neighbourhood of the electrodes which is a direct consequence of a local increase of viscosity. This decrease of velocity tends to diminish the local concentration of ions, which in its turn tends to decrease the local viscosity, and so in the end, with very large fields, the apparent viscosity tends to a saturation value.

The treatment follows the lines first laid down by J. J. Thomson (cf J. J. Thomson and G. P. Thomson, vol. I, chapter IV) for conduction in an ionized gas, but certain essential modifications have to be made, owing to the fact that the ionic mobility is variable.

Let n_1, n_2 be the number per unit volume of positive and negative ions, each carrying charge e, and let α be the coefficient of recombination. Let X be the local strength of the electric field in the direction x, normal to the direction of flow, x being measured from the anode. Then $dX/dx = 4\pi(n_1 - n_2)e$. If the rate of formation of ions be q per unit volume, then

$$\frac{d}{dx}(k_1 n_1 X) = q - \alpha n_1 n_2, \quad \frac{d}{dx}(k_2 n_2 X) = -q + \alpha n_1 n_2,$$

where k_1, k_2 are the respective mobilities of the two kinds of ions.

For simplicity take $k_1 = k_2$ and suppose that the ions are removed so rapidly that the recombination can be neglected, i.e. confine attention to the case of large field strength, for which alone the conclusions here reached will be valid. Then

$$k_1 n_1 X = qx, \quad \text{since } n_1 = 0 \text{ at } x = 0,$$

and

$$k_2 n_2 X = \frac{i}{e} - qx, \quad \text{where } i \text{ is the current density,}$$

$$(n_2 - n_1) = \left(\frac{i}{2qe} - x\right)\frac{2q}{kX}. \tag{i}$$

The local mobility of the ions may be assumed to be inversely as the local viscosity or $k = a/\eta$, where a is constant. Assume, and this is an essential assumption, that the local viscosity is increased by an amount proportional to the number of excess ions, in accordance with the previous argument, or

$$\eta = \eta_0 + b(n_2 - n_1),$$

where η_0 is the viscosity, in the absence of field, and b is a constant. Then from (i),

$$n_2 - n_1 = \frac{2\eta_0 q x'}{aX - 2qbx'}, \quad \text{where } x' = \frac{i}{2qe} - x,$$

whence

$$\frac{dX}{dx'} = \frac{Ax'}{X - Bx'}, \tag{ii}$$

where $A = \dfrac{8\pi e \eta_0 q}{a}$, $B = \dfrac{2qb}{a}$ and A and B are both positive.

The integration of (ii) gives

$$\frac{X^2 - Bx'X - Ax'^2}{\left[\dfrac{2X - (c+B)\,x'}{2X + (c-B)\,x'}\right]^{B/c}} = X_0^2, \quad \text{where } c = \sqrt{(B^2 + 4A)} \tag{iii}$$

within the region $x = 0$ to $x = i/2qe$, i.e. $x' = i/2qe = x_0$ to $x' = 0$. For negative values of x', i.e. $x > x_0$, X has an approximately constant value X_0; in other words, X_0 is the approximately constant value which the field has in the central region.

The field markedly exceeds the value X_0 throughout a layer of thickness x_0 adjacent to the electrode. If I is the saturation current density, then $2dqe = I$, where $2d$ is the separation of the electrodes. Hence $x_0 = (i/I)\,d$. Thus x_0 is proportional to i, or the distance that the layer extends from the electrode increases linearly with the current.

In principle, equation (iii) enables one to find the distribution of X against x for a given voltage V applied to the electrodes. The distribution of X gives the number of excess ions at any point and hence, for an assumed value of b, gives the viscosity as a function of x. From this the time of flow can be found as follows.

Consider any layer distant $\xi = d - x$ from the centre of the channel and let the viscosity be η in this layer. The velocity of this layer along the channel is

$$v = \int_{\xi}^{d} \frac{p}{\eta}\, \xi \, d\xi,$$

where p is the pressure gradient along the channel. If Q_0 is the volume per second delivered at this pressure gradient without electric field, then

$$p = 3Q_0\eta_0/2d^3.$$

Volume per second with electric field $= Q = 2\int_0^d v\,d\xi$

$$= 2\int_0^d\int_\xi^d \frac{p}{\eta}\,\xi\,d\xi\,d\xi.$$

But
$$\eta = \eta_0 + b\,(n_2 - n_1) = \eta_0\left(1 - \frac{B}{A}\frac{dX}{dx}\right),$$

therefore
$$\frac{t}{t_0} = \tfrac{1}{3}d^3 \frac{1}{\displaystyle\int_0^d\int_\xi^d \frac{\xi\,d\xi}{1 - \frac{B}{A}\frac{dX}{dx}}\,d\xi}. \qquad \text{(iv)}$$

Hence a knowledge of the value of B/A, together with the distribution of dX/dx against x, i.e. against ξ, gives the required t/t_0 and hence $(t-t_0)/t_0$ for comparison with observations.

In order to trace the effect of applied voltage on rate of flow, proceed with a numerical example, since the matter is too complicated to permit of a general expression. Take the following estimated values, suggested by acetonitrile:

$$I = 8{\cdot}4\times10^4\,\text{e.s.u./cm.}^2, \qquad k = 0{\cdot}67\,\text{cm./sec./e.s.u.},$$
$$\eta_0 = 3{\cdot}6\times10^{-3}\,\text{c.g.s. units}, \quad d = 10^{-2}\,\text{cm.},$$

whence $A = 1{\cdot}57\times10^8$. To evaluate B an assumption must be made which gives b. Assume that at the electrodes $\eta = 1{\cdot}35\eta_0$; as this is merely to derive a general form for the variations of apparent viscosity with field any reasonable value will do. Since the denominator of the left-hand side of equation (iii) is always less than 1 and near the electrodes X is large compared to X_0, an approximate solution of (iii) is the positive solution of the quadratic

$$X^2 - Bx'X - Ax'^2 = 0,$$

or
$$\frac{X}{x'} = \frac{B + \sqrt{(B^2+4A)}}{2}.$$

But, from (i)
$$B = \frac{2qb}{k\eta} = \frac{X}{x'}\left(1 - \frac{\eta_0}{\eta}\right)$$
$$= \frac{X}{x'}0{\cdot}2593 = 0{\cdot}13(B + \sqrt{[B^2+4A]}) \quad \text{approximately,}$$

whence
$$B = 3{\cdot}8\times10^3.$$

This gives
$$X_0^2\left[\frac{X - 14{\cdot}6\times10^3x'}{X + 10{\cdot}8\times10^3x'}\right]^{0{\cdot}15} = X^2 - 3{\cdot}8\times10^3x'X - 157\times10^6x'^2. \qquad \text{(iii}a\text{)}$$

In order to determine the distribution of field X for a given applied voltage V the constant value of the field X_0 in the central region must be chosen so that the area under the curve of X against x is equal to V.

A value of X_0 being chosen, values of X against x' are obtained from (iii$a$), remembering that $x' = (i/I)\,d = x_0$ at the electrode and $x' = 0$, $X = X_0$ towards the middle of the channel. Consistent values of V, X_0 and X against x' are obtained by successive approximations. Table 5 shows the values of i for given values of V taken as typical of a sample of acetonitrile.

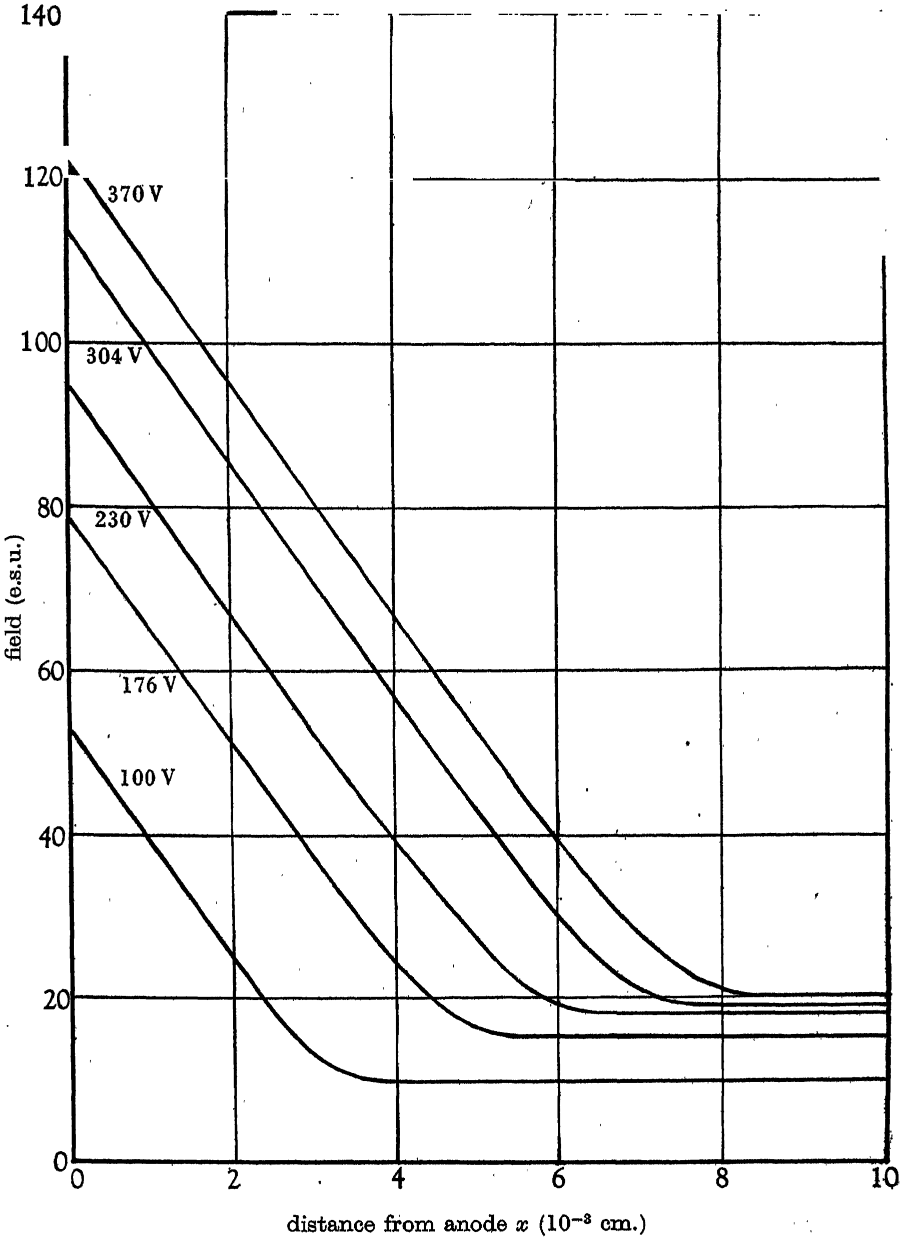


FIGURE 18

TABLE 5

V (V)	i (e.s.u./cm.^{2})	$x_0 = (i/I)\,d$ (cm.)
100	3×10^4	$3{\cdot}57 \times 10^{-3}$
176	4·5	5·35
230	5·4	6·43
304	6·45	7·68
370	7·05	8·40
saturation	8·4	10·00

The method indicated gave the curves of figure 18 for the field distribution at the various applied voltages. Putting appropriate values of X and x' in equation (ii) gives us the value of dX/dx for various values of x (see figure 19). Substituting in equation (iv) and integrating graphically, t/t_0 is obtained as a function of the applied voltage.

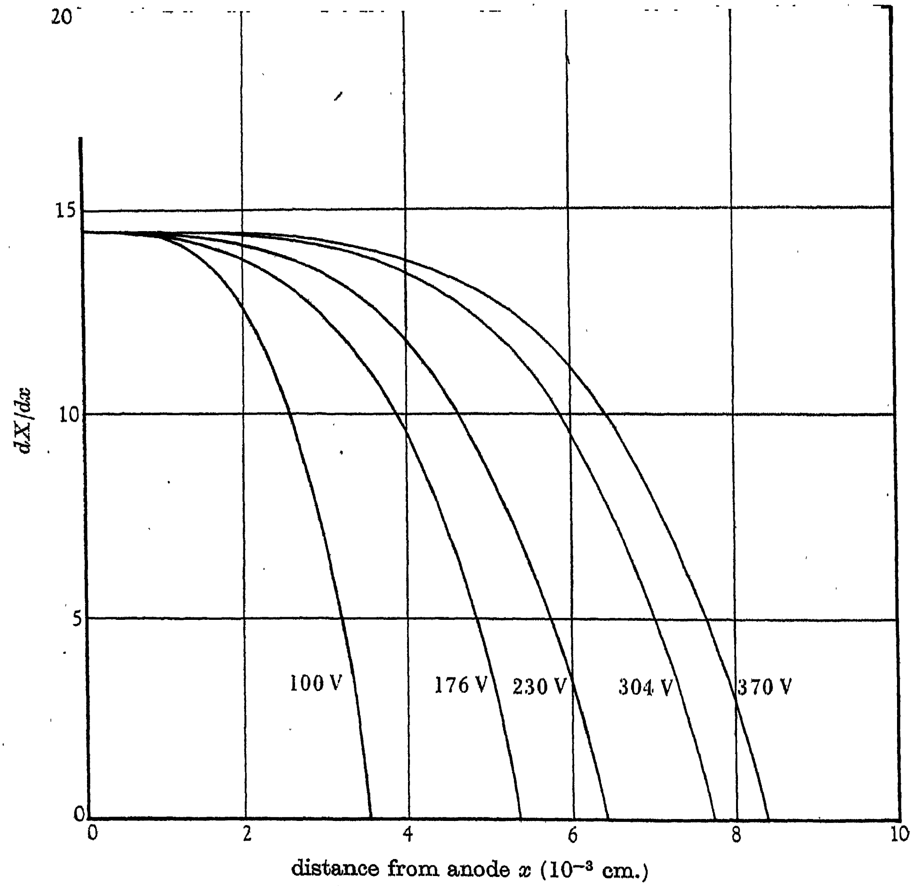


FIGURE 19

The variation of $(t-t_0)/t_0$ with voltages from 100 to 370 is shown in figure 20. This curve clearly shows the tendency for the increase in time of flow to reach a limiting value for high voltages which is so marked a feature of the experimental curves. It will be remembered that the assumptions are valid for high fields only.

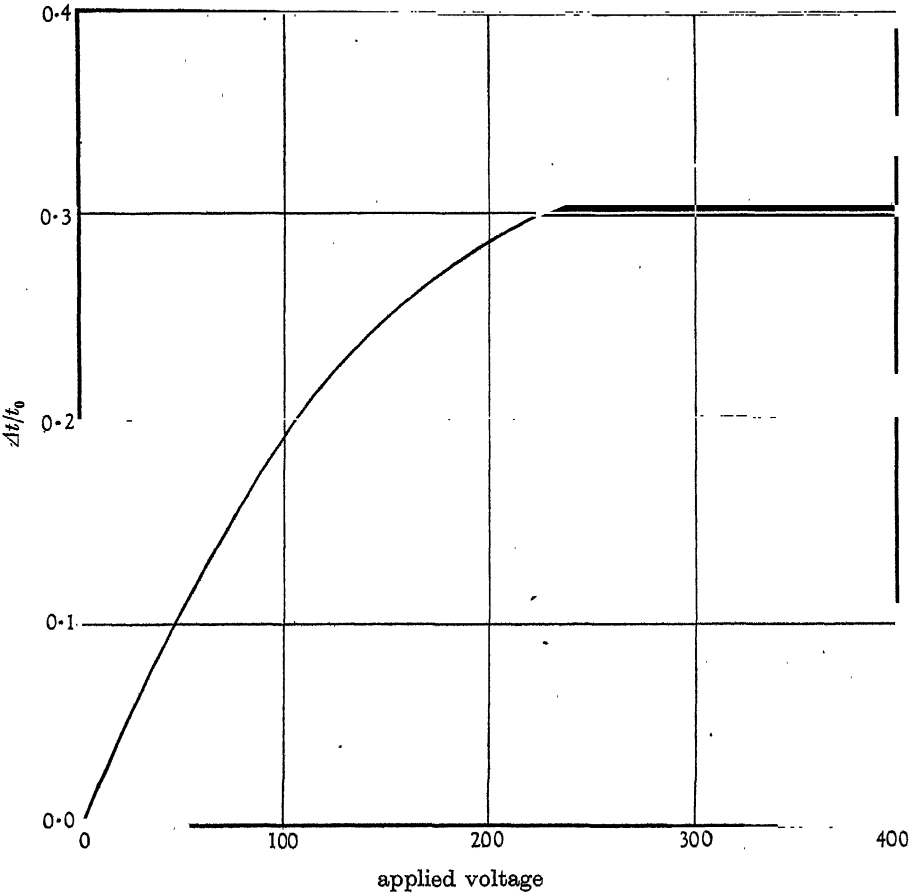


FIGURE 20

The initial portion of the curve in figure 20 for low voltages lies above the values obtained experimentally. The reason for this is that in the above treatment no account has been taken of recombination of the ions. If the fields in the regions near to the electrodes are large, recombination of the swiftly moving ions will be negligible, as has been assumed in deducing the preceding formulae, but if the fields are small the ions will be moving much more slowly and hence some will be lost by recombination. The effect will then be everywhere to reduce the number of excess ions, and hence the viscosity, below the value previously obtained when this effect was neglected. The curve will thus tend to become concave upwards as indicated by the experimental curves. We hope on some future occasion to be able to give the theory of the initial part of the curve, but it is not necessary for the general elucidation of the experimental effects under discussion.

LIQUIDS OBEYING OHM'S LAW

The attainment of a limiting value for the increase in the time of flow is independent of the exact relation between voltage and current. Even if Ohm's law holds at all the voltages used, a limiting value will still be obtained for the time of flow provided that the saturation current which the liquid can carry is not very much larger than the current produced by the largest applied voltage. This follows from the fact that the extent of the layer of high viscosity is governed by the ratio i/I. Had the currents quoted for acetonitrile been proportional to the applied voltage, a curve such as that shown in figure 21 would be obtained as typical of a liquid obeying Ohm's law. The assumption that the increase of viscosity is proportional to the excess of ions of one sign existing locally, combined with the fact that the mobility of the ions is inversely proportional to the local viscosity leads to a relation of the general form found experimentally.

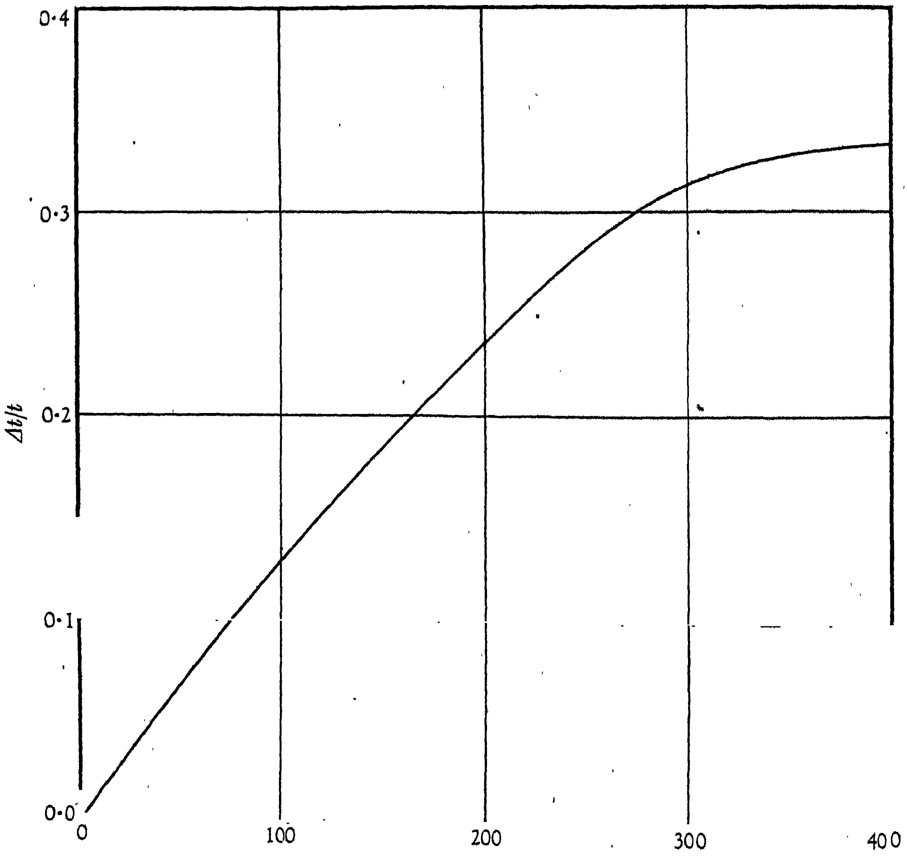


FIGURE 21

EFFECT OF VARYING THE CONDUCTIVITY

The limiting value for the increase in the time of flow is governed almost entirely by the value of the constant $B = 2qb/a$. For any organic liquid both a and b are fixed, but the rate at which the ions are produced, q, is proportional to the electrical conductivity of the sample. The limiting value for the time of flow should therefore be proportional to the conductivity of the liquid, a fact which was observed experimentally for acetonitrile of different degrees of dryness (see figure 4).

EFFECT OF SIZE OF GAP

The extent of the region of increased viscosity measured from the electrode is given by

$$x_0 = (i/I)\,d.$$

Now q, the rate of formation of ions, is given by $q = I/2de$, and since this is constant for identical samples of liquid in two channels of different gaps, the saturation current I must be proportional to the size of the gap ($2d$). Now if the same mean field is applied to the two channels, the same current i will flow and thus x_0, the extent of the region of high viscosity, will be the same. Under these conditions the increase in the time of flow on applying the field will be less in the apparatus having the larger gap, as was observed. The limiting value of the time of flow should, however, be about the same for all gaps, since it is mainly determined by the value of B, which is constant for identical samples of the liquid; with a larger gap, however, a larger mean field must be applied to produce this limiting value.

The effect of the size of the gap on the maximum frequency of the alternating field which gives the full increase in time of flow is as follows. This frequency ν_c must be proportional to X/x_0, where X is the mean field at a distance x_0 from the electrode. If, as a first approximation, the variation of current with applied field be expressed by means of an equation of the form

$$X = \alpha i + \beta i^2,$$

where α and β are positive constants, then

$$\nu_c = A_0 \frac{\alpha i + \beta i^2}{i},$$

since, as has just been seen, x_0 is proportional to i. So

$$\nu_c = A_0(\alpha + \beta i),$$

and obviously decreases as the current decreases.

In the experiments with alternating fields, the applied voltage was approximately constant, and so in the smaller gap apparatus the current, and hence the maximum frequency ν_c, must be bigger than in the apparatus with the larger gap. This is borne out by the experiments on chloroform at 25° C, where the maximum frequency ν_c was 450 sec.$^{-1}$ for a gap of 0·20 mm. and 350 sec.$^{-1}$ for a gap of 0·28 mm.

IS THERE A GENUINE EFFECT OF ELECTRIC FIELD ON VISCOSITY?

We have established the conditions under which large apparent increases of the viscosity of conducting liquids may take place in an electric field, and have shown that these effects occur in polar liquids only when there are large concentrations of ions in the neighbourhood of the electrodes.

The increase of viscosity observed is thus, in a sense, spurious, since the part of the liquid in which the increase of viscosity occurs is not normal but contains artificially produced clusters.

This spurious effect can be avoided by using a rapidly alternating field. This avoids concentrations of ions near the electrodes which not only gets rid of clusters but also ensures that the field shall be uniform.

In our experiments with chloroform the mean applied potential difference was about 10 kV/cm., and the total time of flow was about 300 sec. With care a difference of time of 0·2 sec. could have been detected, which means a change of viscosity of 1 part in 1500. However, at the time when the experiments on the effect of variation of frequency were being carried out a general survey only of the effect was in view, and the accuracy was not pressed to the limit. For chloroform at 15° C, $E/kT = 2{\cdot}15$, and the *mean* field in the experiments was 10 kV/cm. E thus has the value $8{\cdot}5 \times 10^{-14}$ and μ is about 1×10^{-18}. This leads, on the considerations given at the beginning of this paper, to the value $d\eta/\eta = 8 \times 10^{-4}$, which is just on the limits of detection unless special precautions are taken. It should certainly not be beyond ultimate detection, since both the field and the time of flow can be increased.

In general, the way in which a true effect is to be sought has been established by these experiments. An alternating field must be used, whose frequency is well in excess of $uV/2D^2$, where u is the mobility, V is the applied potential and D is the gap between the electrodes. If D were reduced to 0·01 cm. and the gradient V/D taken as 40 kV/cm., say, this means that a frequency well in excess of 1000 should be used. At the same time the frequency should not be high enough to produce effects involving the relaxation time. Luckily such effects do not set in below a frequency of about 10^7 cyc./sec. Frequencies in the range 10^4–10^6 cyc./sec. should, therefore, be suitable. Our experiments have suggested to us various ways in which the apparatus could be modified to increase somewhat the sensitiveness of the method. There is, therefore, every prospect that an effect of the order to be anticipated can be measured under precise and significant conditions, more especially as with an alternating field the potential gradient can probably be considerably increased without danger of electric breakdown.

REFERENCES

Alcock, E. D. 1936 *Physics*, 7, 126.
Andrade, E. N. da C. 1934 *Phil. Mag.* **17**, 497, 698.
Andrade, E. N. da C. 1940 *Proc. Phys. Soc. Lond.* **52**, 748.
Andrade, E. N. da C. & Dodd, C. 1939*a* *Nature*, **143**, 26.
Andrade, E. N. da C. & Dodd, C. 1939*b* *Nature*, **144**, 117.
Björnståhl, Y. & Snellman, O. 1935 *Physics*, **6**, 257.

Björnståhl, Y. & Snellman, O. 1937 *Kolloidzschr.* **78**, 258.
Dobinski, S. 1935 *Phys. Z.* **36**, 509.
Duff, W. 1896 *Phys. Rev.* **4**, 23.
Ercolini, G. 1903 *Nuovo Cim.* **5**, 249.
Eyring, H. 1936 *J. Chem. Phys.* **4**, 283.
Herzog, R. O., Kudar, H. & Paersch, E. 1933 *Naturwissenschaften*, **21**, 662.
Herzog, R. O., Kudar, H. & Paersch, E. 1934 *Phys. Z.* **35**, 446.
Jaffé, G. 1908 *Ann. Phys., Lpz.*, **25**, 257.
Jaffé, G. 1910 *Ann. Phys., Lpz.*, **32**, 148.
Kimura, O. 1937 *Bull. Chem. Soc. Japan*, **12**, 147.
König, W. 1885 *Ann. Phys., Lpz.*, **25**, 618.
Pacher, G. & Finazzi, L. 1900 *Nuovo Cim.* **11**, 290.
Pochettino, A. 1903 *R.C. Accad. Lincei*, **12**, II, 363.
Quincke, G. 1897 *Ann. Phys., Lpz.*, **62**, 1.
Schweidler, E. von 1901*a* *Ann. Phys., Lpz.*, **4**, 307.
Schweidler, E. von 1901*b* *Ann. Phys., Lpz.*, **5**, 483.
Schweidler, E. von 1904 *S.B. Akad. Wiss. Wien*, **13**, 881.
Sokolov, P. & Sosinski, S. 1935 *C.R. Acad. Sci. U.R.S.S.* **4**, 135.
Sokolov, P. & Sosinski, S. 1936 *Acta Physicochim. U.R.S.S.* **5**, 691.
Sokolov, P. & Sosinski, S. 1939 *Nature*, **144**, 117.
Thomson, J. J. & Thomson, G. P. 1928 *Conductivity of electricity through gases*, **1**. Cambridge University Press.
van der Bijl, H. J. 1912 *Ann. Phys., Lpz.*, **39**, 170.

The thermochemistry of carbon: valence states, heats of sublimation and energies of linkage

BY L. H. LONG, D.I.C., PH.D., AND R. G. W. NORRISH, F.R.S.
Laboratory of Physical Chemistry, Cambridge

(*Received* 10 *January* 1946—*Read* 2 *May* 1946)

The controversy which exists at the present time between the figures 125 and 170 kcal./g.-atom for the latent heat of sublimation of carbon into monatomic vapour in the ground state originates largely from the neglect to take into consideration the energy required to raise the carbon atoms from the ground (^3P) state to the lowest tetravalent (^5S) electronic configuration corresponding to that in which it is normally found in chemical combination. Consideration of the energies of removal of a hydrogen atom from the methane and ethane molecules and of the energies of reorganization of the resulting radicals leads to the figure $190 \pm$ about 10 kcal. for L_2, the heat of sublimation into free atoms in the ^5S state. This in turn leads to a satisfactory and unambiguous assignment of values to bond energies (as distinct from dissociation energies) which can now be expressed with an uncertainty of not more than a few kcal. In the light of the valency distinction there remains no sound evidence to maintain the higher value put forward for L_1, and 125 kcal. is unquestionably of the right order. There are strong indications that an earlier estimate of 100 kcal. for the energy level of the ^5S state above the ^3P (ground) state is about 50 % in excess of the true value.

The necessity for establishing this branch of thermochemistry on a sound theoretical and experimental footing has long been a very obvious need. The scheme here suggested reconciles points hitherto in apparent conflict, and brings virtually all established experimental knowledge into alignment.

1. INTRODUCTION

The dual purpose of this paper is to review critically the divergent values published for the heat of sublimation of carbon and the strengths assigned to certain chemical bonds involving carbon, including the reported variations in bond strength with the length of carbon chain. The two subjects are interdependent, and observations we wish to put forward with respect to the one throw a certain amount of light on the other.

Two prerequisites to the clear understanding of the thermochemistry of carbon and, indeed, of thermochemistry in general, are:

(*a*) a careful distinction between the different natures of bond energies and energies of dissociation;

(*b*) a differentiation between the fundamentally different energy levels of each individual element in its various states of valency (being two and four for carbon).

Both distinctions have been ignored by many in the past, and this has been largely responsible for the uncertain and conflicting bond-energy values which have been a feature of the thermochemistry of carbon from its beginning. The stress on the importance of these differences is not new in either instance, but a systematic treatment of this field in the light of these essential distinctions has not been undertaken before, and a reconciliation of many seemingly conflicting facts is here proposed.

2. MEASUREMENTS AND CALCULATIONS

It will help to clarify our arguments if we briefly review the methods which are at present available for the determination of the energies associated with chemical linkages. The following methods have been employed, of which the first two have had the widest application:

(*a*) *Thermochemical.* The heats of reaction of a number of different types of chemical processes have been measured and employed in the calculation of heats of formation and bond energies. For this purpose the heats of combustion have been found in general to be the most accessible. In the case of binary compounds with only one type of linkage, these can be combined directly with the heats of combustion of the component elements in their normal states, together with the heats of formation of the elements in these states from their monatomic vapours, to give the required bond energies. For example, if the heat of combustion of hydrogen sulphide to sulphur dioxide and water is accepted to be 134·5 kcal./mol., the energy of the S—H bond could be calculated in the following manner:

$$
\begin{array}{rcll}
 & & & \text{kcal.} \\
H_2S + 1\tfrac{1}{2}O_2 & = & SO_2 + H_2O\ (\text{liq.}) & +134{\cdot}5 \\
SO_2 & = & S\ (\text{rhombic}) + O_2 & -\ \ 70{\cdot}91 \\
S\ (\text{rhombic}) & = & S(^3P) & -\ \ 66{\cdot}3 \\
H_2O\ (\text{liq.}) & = & H_2 + \tfrac{1}{2}O_2 & -\ \ 68{\cdot}32 \\
H_2 & = & 2H & -104{\cdot}1 \\
\hline
(\text{By addition})\ \ H_2S & = & S(^3P) + 2H & -175{\cdot}1
\end{array}
$$

The last process involves only the rupture of two S—H bonds. Thus the average S—H bond energy is 87·6 kcal.

The purpose of providing this simple example is to emphasize that, in the case of polyvalent elements, the value provided by thermochemical data is the average value of the bond energy (E) and not the dissociation energy (D) of any specific linkage.

Heats of combustion are frequently expressed with uncertainties of 3 parts or less in 10,000, but not all thermochemical quantities are known with anything approaching this precision. The most notable example of this is the energy difference between solid carbon and its monatomic vapour in the ground state, that is, the latent heat of sublimation of carbon. Opinions are divided between two values differing by about 30 %.

(*b*) *Spectroscopic.* It is frequently possible to obtain information about energies of dissociation from band spectra. The various ways of doing this are reviewed elsewhere (Herzberg 1937). Where a convergence limit can be observed directly, the required energy may be obtained with considerable accuracy, that is, to within a few tenths of 1 %. Accurate values may also be furnished by predissociation limits if the potential energy curves for the products satisfy certain conditions. Otherwise the available methods supply in general only upper limits for dissociation energies.

To obtain energies of dissociation into atoms or radicals in the ground state it is essential to determine the energy levels of the products, if they are not originally formed in that state. Also, for molecules which might dissociate in more than one way, wrong conclusions may be drawn unless evidence for the primary step is obtained by an examination of the products of photodecomposition at the relevant wave-length.

In some cases it has not been possible to decide conclusively between several possible values. This is so in the case of carbon monoxide where it is not known with certainty if the predissociation reported at 9·61 eV is real, accidental, or merely ordinary perturbation. If $D_0(CO)$ could be unambiguously determined, it would be possible, since the other requisite thermochemical quantities are accurately known, to fix the value for the latent heat of sublimation of carbon.

(*c*) *Reaction kinetics.* From a knowledge of the kinetics of reactions involving free atoms or radicals, and a determination of the activation energies of the reactions concerned, it is possible to derive information on dissociation energies. This method has been recently applied, for example, to estimate the energies of removal of the first hydrogen atom in methane (Andersen, Kistiakowsky & Van Artsdalen 1942; Andersen & Kistiakowsky 1943; Kistiakowsky & Van Artsdalen 1944) and ethane (Andersen & Van Artsdalen 1944).

(*d*) *Electron impact critical potentials.* By making certain rather questionable assumptions, Stevenson (1942) has attempted to calculate $D(CH_3—H)$ and $D(C_2H_5—H)$ from the critical potentials observed in electron impact experiments. This method has also been successfully employed by others to diatomic molecules.

The results agree reasonably well with those obtained by other methods, allowance being made for the inherent uncertainty of several percent.

(*e*) *Other methods*. In the case of reversible reactions, equilibrium data have been much used to calculate energies of dissociation by means of the van 't Hoff isochore. It is also possible to calculate D for a neutral molecule by combining D for an ionized state with ionization energy data. For a few very simple atomic linkages, D has been calculated by exhaustive wave-mechanical or other theoretical treatment, as in the case of the H—H bond (James & Coolidge 1933; Cook 1945). Finally, there is a semi-empirical relationship deduced by Sutherland (1938, 1940), for the interdependence of interatomic distance, force constant and dissociation energy, from which a rough value of the latter may be derived if sufficient data are available. These methods do not closely concern us here and need not be discussed further.

In order to avoid confusing the bond energy E with the energy of dissociation D, consideration must be given to the precise quantity provided by the method in question. Only for diatomic molecules can D and E sometimes be equated. Although in the case of polyatomic molecules the two quantities may on occasions approximate quite well, it needs to be emphasized that they are not identical and frequently differ by large amounts. It is seen that D depends on the states of the products, whereas E, like the bond length, has a fixed value.

3. THE TWO HEATS OF SUBLIMATION OF CARBON

It is of primary importance to remember, when employing heats of formation to calculate the sum of the bond energies in molecules containing di- and tetravalent carbon, that the molecules must be regarded in the two cases as having been built up from carbon atoms initially possessing fundamentally different energy levels (Norrish 1934). The ground state of the carbon atom corresponding to a valency of two is s^2p^2, ^3P, the lowest corresponding to a valency of four being sp^3, ^5S, which is at a considerably higher energy level. The energy change in the $^5S \rightarrow {}^3P$ transformation has not been directly determined for C, although the corresponding changes have been observed for N^+ and O^{++}. By an extrapolation method, the value for C has been estimated by Bacher & Goudsmit (1934) to be about 4·3 eV or 100 kcal., but, more recently, Ufford (1938) has placed it at about 73 kcal. If the latent heats of sublimation of graphite into monatomic vapours consisting entirely of atoms in the ^3P and ^5S states be designated by L_1 and L_2 respectively, then, assuming the former figure quoted to be correct, $L_2 - L_1 = 100$ kcal./mol. However, we believe the difference between L_2 and L_1 to be considerably less than this amount for reasons which will be discussed later.

Neither L_1 nor L_2 is known with certainty. Whereas some authors support a value for L_1 of about 125 kcal./g.-atom (Herzberg 1937, 1939, 1942; Goldfinger & Jeunehomme 1936), others advocate a much higher figure around 170 kcal. (Gaydon & Penney 1945; Kynch & Penney 1941; Baughan 1941). Clearly, since carbon in most of its compounds is tetravalent, the figure with which thermochemical data are normally concerned in the first instance is L_2 rather than L_1. This point has escaped many investigators, including Pauling (1939), whose table of bond energies consequently contains values deficient in the case of links involving carbon. For example, if the average bond energy in methane were known, L_2 could be calculated from the following set of equations:

$$\begin{array}{rcll} & & & \text{kcal.} \\ CH_4 & = & C\ (\text{graphite}) + 2H_2 & -\ 17{\cdot}87 \\ 2H_2 & = & 4H & -208{\cdot}2 \\ C\ (\text{graphite}) & = & C(^5S) & -L_2 \\ \hline CH_4 & = & C(^5S) + 4H & -(L_2 + 226{\cdot}1) \end{array}$$

Thus $L_2 + 226{\cdot}1$ kcal. could be placed equal to four times the average C—H bond energy and L_2 evaluated directly.

On the other hand, if one considers the total energy required to rupture the four bonds in methane by successive steps, one would here be primarily concerned with L_1 if the carbon were able to revert to the divalent state with the rupture of the second bond. That the carbon in CH_2 is indeed in the divalent state is indicated by the experimental work of Pearson, Purcell & Saigh (1938) and Barrow, Pearson & Purcell (1939). If the energy absorbed by the four steps were a, b, c and d respectively, then

$$CH_4 = C(^3P) + 4H - (a+b+c+d).$$

The value of a has been determined to be 102 kcal. (Andersen & Kistiakowsky 1943; Kistiakowsky & Van Artsdalen 1944). One would expect b to have the same order of magnitude but for the fact that during the step $CH_3 \rightarrow CH_2 + H - b$, the carbon changes to the divalent state with the liberation of energy. This energy may be regarded as being divided into three parts: one part is liberated during the removal of the hydrogen atom, which means that b will be smaller than the mean bond energy in free methyl, whereas the other two parts are absorbed in the weakening of the two remaining bonds. The value of d, the energy of removal of the last hydrogen atom, is known to be 80 kcal., having been recalculated by Herzberg (1939) from published spectroscopic data (Shidei 1936). There is nothing to support the view of Berriman & Clark (1938) that d is $\frac{1}{4}(L_2 - L_1)$ less than the mean bond energy in methane. c would be expected to have about the same magnitude as d, and a theoretical treatment by Voge (1936) indicates that the difference is only about 0·05 eV. Thus, in the calculation of L_1, c could be placed equal to d as a first approximation. If were known, therefore, L_1 could be determined from the two foregoing equations and the relationship

$$C(^5S) = C(^3P) + L_2 - L_1.$$

Although b is not known, it has long been recognized that its value is probably considerably less than a (Mecke 1930). This is questioned by Baughan (1941), however, who puts forward the following two arguments in support of the higher value 170 kcal. for L_1:

(*a*) The first is based on the work of Voge (1936), who calculates that the energy required for the successive removal of hydrogen atoms in methane does not vary by any large amount. Voge's conclusions, however, rest on the quantities 17·0 and 4·0 eV for the atomic heats of formation of CH_4 and CH respectively. Considerable uncertainty must be attached to the former figure, and the latter is now known to be over 0·5 eV less than the value employed (Herzberg 1939). Furthermore, it is now known that the process $CH_4 \rightarrow CH_3 + H$ requires 101 kcal. or 4·38 eV at absolute zero (Kistiakowsky & Van Artsdalen 1944), and, combining this value with Voge's equations (20) and (22), a new value of 1·16 eV for his integral $N_{s\sigma}$ is obtained, which in turn leads to the estimates 14·25 and 9·87 eV respectively for the energies of formation of CH_4 and CH_3 from free atoms. These figures lead to a low value for L_1, even if the fact that they include zero-point energy be neglected. Thus Voge's calculations are not compatible with the value 170 kcal. for L_1.

(*b*) The second argument rests on the observation that the energy exchange in the successive replacement of hydrogen atoms in methane by halogen atoms is approximately constant for any one halogen. This does not seem to us, however, to be a relevant argument, since in every case the carbon atom remains saturated and there are no free valencies or divalent state to consider as in the case in question.

We will now examine the arguments of Kynch & Penney (1941) who also support this value of 170 kcal. for L_1, claiming agreement for several independent methods as follows:

	method	L kcal.
(i)	spectroscopic determination of $D_0(CO)$	168·8
(ii)	rate of sublimation of carbon *in vacuo*	176[*]
(iii)	heat of reaction $2CH_4 \rightarrow C_2H_6 + H_2$	170
(iv)	Birge-Sponer extrapolation for CO	~175
(v)	absorption-band limits of (*a*) butadiene, (*b*) hexatriene, (*c*) benzene	168–173

On close examination, however, it becomes apparent that some of these methods provide a value for L_1 and some for L_2.

Methods (i) and (iv) lead to L_1, but the value selected for $D_0(CO)$ is only one of several possible values, others of which, for example, lead to $L_1 = 125$ or 108·5 kcal. Method (iv), as has been pointed out by Herzberg (1942), usually provides values 20–40 % too high for long extrapolations such as are involved here, thus supporting $L_1 = 125$ kcal. rather than $L_1 = 170$ kcal.

Method (ii) is due to Marshall & Norton (1933) who determined the rate of loss in weight of carbon at high temperatures *in vacuo*, and calculated that the conversion of solid carbon to monatomic vapour requires 177 kcal./g.-atom. Since solid carbon is in the tetravalent state, it cannot be assumed that the vapour is produced directly in the ^3P state, though doubtless some atoms quickly revert to this. Only, therefore, in a method which permitted the establishment of equilibrium conditions could one feel sure that, allowing for the additional complication of the presence of C_2

[*] Misquoted by Kynch & Penney. The experimental figure is 177 kcal.

molecules, a measure of L_1 was being obtained. Here is a dynamic method, which, if the carbon vapour were initially produced in the ^5S state, would provide a value for L_2. That the vapour, or at least the major part of it, is not initially produced in the ^3P state is indicated by the wide disagreement in values provided by the equilibrium and dynamic methods (§ 5). The manner in which carbon vapour leaves a solid surface is not known, but several suggestions have been made (Herzberg, Herzfeld & Teller 1937; Herzberg 1937; Kynch & Penney 1941). If Herzberg's considerations of a 'potential hill' apply, then the dynamic method should provide a value between L_1 and L_2. It must be concluded, therefore, that in the absence of a precise knowledge of the mechanism of sublimation of carbon, the method of Marshall & Norton cannot provide an experimental value for L_1, but might be more rightly regarded as furnishing a lower limit for L_2.

Method (iii) provides an approximation to L_2. Since L_2 is numerically equal to twice the energy of the single C—C link in solid carbon (§ 4), and since this link is rather stronger than the C—C link in ethane (§ 4), the value 88·2 kcal. calculated for the latter link (Baughan & Polanyi 1940; Baughan 1941) leads to a value for L_2 nearer 180 than 170 kcal. But for this calculation Baughan tacitly identifies bond energies with dissociation energies as will be discussed in greater detail in the next section.

Likewise, when discussing bond energies, Kynch & Penney fail to realize that the figures they quote refer to L_2 and not to L_1. If method (v) is valid, it determines L_2; but, without stating their reasons, Mulliken & Rieke (1942) consider it to provide no basis for deciding upon the true value of the heat of sublimation of carbon.

In conclusion, therefore, the apparent agreement obtained by Kynch & Penney is seen to be unreal and misleading. Inasmuch as the higher value (170 kcal.) is based on arguments leading to L_2, there are no grounds in these considerations of Kynch & Penney for regarding this quantity as incompatible with $L_1 = 125$ kcal., the value these authors are trying to refute in their belief that $D_0(\mathrm{CO}) = 11{\cdot}1$ eV and $L_1 = 170$ kcal. Nevertheless, though the figure 170 kcal. may be regarded as not altogether an unreasonable estimate for L_2, in our opinion it still needs some modification to comply with published experimental data (§ 5).

4. VARIATION OF BOND STRENGTHS IN THE NORMAL PARAFFIN SERIES

Until 1933, published values for the heats of combustion of the series of straight-chained paraffins showed no irregularities outside the limits of experimental error. There was thus no reason to believe the C—C and C—H bond strengths were other than constant (Sidgwick & Bowen 1931). However, more accurate measurement has indicated a deviation in bond strength towards increased stability of the first five members of the series (Rossini 1934*b*), the deviation decreasing from methane to *n*-pentane.

Van Artsdalen (1942) has calculated the values of

$$D(\mathrm{CH_3—CH_3}) \quad \text{and} \quad D(\mathrm{C_2H_5—C_2H_5}),$$

obtaining a difference of about 3 kcal., to which he attaches a qualitative rather than quantitative significance in view of the probable errors. A similar decrease with increasing chain length has been reported by Stevenson (1942). Unfortunately, Van Artsdalen, in common with Baughan (1941), confuses the issue somewhat by identifying the energy of dissociation with the bond strength. In the case of the C—C link in C_2H_6, for example, the dissociation energy differs from the bond strength by six times the difference between the C—H bond strengths in CH_3 and C_2H_6, which may amount to several kcal. The difference between

$$D(CH_3—CH_3) \quad \text{and} \quad D(C_2H_5—C_2H_5)$$

is reduced to about 1 kcal. when the more recent figure for $D(C_2H_5—H)$ (Andersen & Van Artsdalen 1944) is employed in Van Artsdalen's calculations. Let us now proceed to investigate the actual variation with chain length of the average values for E(C—C) and E(C—H).

In both diamond and graphite the carbon is in the tetravalent state, and the sublimation of either solid form into monatomic vapour involves the energy of rupture of two valency links per carbon atom (plus a small RT term which will be neglected). Thus the average strength of the C—C valency link is $\frac{1}{2}L_2$. Since the difference in the heats of combustion of diamond and graphite is negligibly small, being only a few tenths of 1 kcal./g.-atom (Prosen, Jessup & Rossini 1944), the heats of sublimation are almost identical, and the C—C bond strength in diamond may be equated to $\frac{1}{2}L_2$ without appreciable error.

The aliphatic C—C bond is known to be slightly weaker than that of diamond, varying values for this difference having been suggested from time to time. The order of this amount is given by some of the later papers dealing with this point at 4 kcal. (Lasareff 1935; Schmid 1936). This figure, however, can be obtained with greater accuracy from the heats of formation of the paraffin series. For this purpose the difference in strength between the central C—C bond in the long-chained paraffins and the C—C bond in solid carbon will be equated to x kcal. The strength of the aliphatic bond then becomes $\frac{1}{2}L_2 - x$. The strength of the H—H bond is required for the calculation of absolute values, and this is known accurately. Herzberg (1939) gives it as 103·2 kcal. at 0° K, which is equivalent to 104·1 kcal. at 25° C and constant pressure when the appropriate adjustment of 0·94 kcal. (Giauque 1930) is made.

One cannot, of course, calculate absolute values for the average bond energies while L_2 remains undetermined, but by combining the foregoing values with the heats of formation of the paraffins it is possible to obtain relative values, the variation being independent of L_2. The heats of formation of the lower homologues from methane to n-pentane have been calculated (Rossini 1940) direct from the heats of combustion (Rossini 1931, 1934a). For the higher members of the series, from n-hexane to n-dodecane, the heats of formation have been calculated from the heats of combustion in the liquid state (Jessup 1937) and the latent heats of vaporization (Holcomb & Brown 1942). The accuracy will not be quite as high for the liquid paraffins, as the latent heats may be in error by a small percentage, but the error thereby introduced in the bond energies will be very small as the latent heats are small compared with the heats of combustion.

The heats of formation of the higher members of the series from *n*-hexane upwards show regular increments of 5·45 kcal. per CH_2 group, that is, the heat of formation of the CH_2 grouping is equal to this amount. Removing a CH_2 group from the middle of a chain to leave a molecule containing one carbon atom fewer eliminates one C—C and two C—H bonds. The sum of these bond energies is equal to the energy required to convert the CH_2 grouping into atoms, which may be simply calculated:

$$\begin{array}{rll} & & \text{kcal.} \\ CH_2\ \text{(grouping)} = & C\ \text{(graphite)} + H_2 - & 5{\cdot}45 \\ C\ \text{(graphite)} = & C(^5S) & -L_2 \\ H_2 = & 2H & -104{\cdot}1 \\ \hline CH_2\ \text{(grouping)} = & C(^5S) + 2H & -(L_2 + 109{\cdot}55) \end{array}$$

Hence $E(C—C) + 2E(C—H) = L_2 + 109{\cdot}55$ kcal. But $E(C—C) = \frac{1}{2}L_2 - x$, from which it follows

$$E(C—H) = \tfrac{1}{4}L_2 + \tfrac{1}{2}x + 54{\cdot}775\ \text{kcal.}$$

This value can be compared with the corresponding values for the paraffins themselves. The following set of equations is known for the case of methane (the third place of decimals, although of no significance in the actual bond-energy values, is necessary for the accurate computation of small differences):

$$\begin{array}{rll} & & \text{kcal.} \\ CH_4 = & C\ \text{(graphite)} + 2H_2 - & 17{\cdot}865 \\ C\ \text{(graphite)} = & C(^5S) & -L_2 \\ 2H_2 = & 4H & -208{\cdot}2 \\ \hline CH_4 = & C(^5S) + 4H & -(L_2 + 226{\cdot}065) \end{array}$$

Thus for methane, $E(C—H) = \frac{1}{4}L_2 + 56{\cdot}516$ kcal.

For the other paraffins allowance must be made for the C—C bonds. If the C—C bond energy, as a first approximation, is assumed constant at $\frac{1}{2}L_2 - x$, the series of values given in the second column of table 1 is obtained. The difference between the average values of E(C—H) and that for the CH_2 group, when multiplied by the number of C—H bonds in the molecule, gives δ, the total deviation from linearity of the heats of formation of the various hydrocarbons as shown in the third column. Since the deviation for *n*-hexane and the higher members is zero, the value of x is seen to be 0·93 kcal. Still assuming the C—C bond to be constant, the average energies of the C—H bond are compared in table 2. The figures in the third column are derived from the heats of formation at absolute zero as calculated by Pitzer (1940), and are given for comparison. The superscript in the symbol L_2^0 refers to the absolute temperature. (The heat of vaporization data for the liquid homologues have not been obtained from the same source in columns 2 and 3.)

TABLE 1

substance	average E(C—H) at 25° C kcal.	δ kcal.
CH_4	$\frac{1}{4}L_2+56{\cdot}516+0{\cdot}000x$	$6{\cdot}96-2x$
C_2H_6	$\frac{1}{4}L_2+55{\cdot}415+0{\cdot}167x$	$3{\cdot}84-2x$
C_3H_8	$\frac{1}{4}L_2+55{\cdot}144+0{\cdot}250x$	$2{\cdot}95-2x$
n-C_4H_{10}	$\frac{1}{4}L_2+55{\cdot}022+0{\cdot}300x$	$2{\cdot}47-2x$
n-C_5H_{12}	$\frac{1}{4}L_2+54{\cdot}945+0{\cdot}333x$	$2{\cdot}04-2x$
n-C_6H_{14}	$\frac{1}{4}L_2+54{\cdot}908+0{\cdot}357x$	$1{\cdot}86-2x$
n-C_8H_{18}	$\frac{1}{4}L_2+54{\cdot}878+0{\cdot}389x$	$1{\cdot}86-2x$
n-$C_{12}H_{26}$	$\frac{1}{4}L_2+54{\cdot}847+0{\cdot}423x$	$1{\cdot}86-2x$
CH_2 grouping	$\frac{1}{4}L_2+54{\cdot}775+0{\cdot}500x$	—

TABLE 2

substance	average E(C—H) at 25° C kcal.	average E(C—H) at 0° K kcal.
CH_4	$\frac{1}{4}L_2+56{\cdot}52$	$\frac{1}{4}L_2^0+55{\cdot}57$
C_2H_6	$\frac{1}{4}L_2+55{\cdot}57$	$\frac{1}{4}L_2^0+54{\cdot}50$
C_3H_8	$\frac{1}{4}L_2+55{\cdot}38$	$\frac{1}{4}L_2^0+54{\cdot}26$
n-C_4H_{10}	$\frac{1}{4}L_2+55{\cdot}30$	$\frac{1}{4}L_2^0+54{\cdot}20$
n-C_5H_{12}	$\frac{1}{4}L_2+55{\cdot}25$	$\frac{1}{4}L_2^0+54{\cdot}16$
n-C_6H_{14}	$\frac{1}{4}L_2+55{\cdot}24$	$\frac{1}{4}L_2^0+54{\cdot}15$
n-C_8H_{18}	$\frac{1}{4}L_2+55{\cdot}24$	$\frac{1}{4}L_2^0+54{\cdot}14$
n-$C_{12}H_{26}$	$\frac{1}{4}L_2+55{\cdot}24$	$\frac{1}{4}L_2^0+54{\cdot}14$
CH_2 grouping	$\frac{1}{4}L_2+55{\cdot}24$	$\frac{1}{4}L_2^0+54{\cdot}12$

For the higher members of the series the following relationship holds at 25° C:

$$E(\text{C—C}) = 2E(\text{C—H}) - 111{\cdot}4\,\text{kcal.}$$

This relationship is independent of the heat of sublimation of carbon, and depends solely on accurately known heats of combustion and on heats of vaporization.

The foregoing tables are based on the assumption that the C—C bond strength is constant. It is far more reasonable to suppose that the C—C bonds also make a contribution to the deviation from linearity of the lower homologues. The proportion of the deviation due to each type of bond cannot be assessed with certainty. It seems reasonable to suppose that the variation in bond energy rests mainly in the contribution due to the carbon atoms, since the proportion of primary and secondary carbon atoms varies from molecule to molecule. This would lead one to expect that the variation in the average strengths of the C—C bonds (involving two carbon atoms) would be about twice that of the C—H bonds. On this assumption the mean strengths of the two bond types for the series are compared in table 3. The maximum variation in the average bond strength is seen to be about 1·3 kcal. for the C—H bond and 0·5 kcal. for the C—C bond. For the whole series the C—H bond shows appreciable deviation only in the case of methane. The deviation for ethane is very small, and for the other homologues is negligible (less than 0·1 kcal.). Thus the postulate of the additivity of bond strengths holds very accurately for the normal paraffins with the exceptions of methane and, to a lesser extent, ethane. The difference in E(C—H) in CH_4 and C_2H_6 is seen to be about 1 kcal., and the C—C link in C_2H_6 roughly 0·5 kcal. weaker than in solid carbon. The variation of D(C—H) for individual primary, secondary and tertiary hydrogen atoms is dealt with elsewhere (Brackett 1928; Eyring 1932; Conn, Kistiakowsky & Smith 1939; Smith & Taylor 1939).

TABLE 3

substance	average E(C—C) at 25° C kcal.	average E(C—H) at 25° C kcal.
CH_4	—	$\frac{1}{4}L_2 + 56{\cdot}52$
C_2H_6	$\frac{1}{2}L_2 - 0{\cdot}44$	$\frac{1}{4}L_2 + 55{\cdot}49$
C_3H_8	$\frac{1}{2}L_2 - 0{\cdot}74$	$\frac{1}{4}L_2 + 55{\cdot}33$
n-C_4H_{10}	$\frac{1}{2}L_2 - 0{\cdot}86$	$\frac{1}{4}L_2 + 55{\cdot}28$
n-C_5H_{12}	$\frac{1}{2}L_2 - 0{\cdot}92$	$\frac{1}{4}L_2 + 55{\cdot}25$
n-C_6H_{14}, etc.	$\frac{1}{2}L_2 - 0{\cdot}93$	$\frac{1}{4}L_2 + 55{\cdot}24$

Shortly before going to press there has been published a revised set of values for the heats of formation of the paraffins at 25° C (Prosen & Rossini 1945) and at 0° K (Prosen, Pitzer & Rossini 1945). The most surprising feature about these new figures is that the first five members of the series no longer all deviate towards greater stability as originally indicated (Rossini 1934*b*). Methane is still given as the most stable, but n-pentane, n-butane and propane deviate towards decreased stability, the bond energies reaching a minimum in the case of propane. This lessening of stability is very slight but seems to exceed the limits of experimental error. With these figures the value of x required is 2·7 kcal. and the maximum variation in the average bond strengths is smaller, being about 0·7 kcal. for the C—H bond and less than 0·1 kcal. for the C—C bond. The value of E(C—H) is about 0·7 kcal. stronger for methane than for ethane. If accepted, this revision would not affect materially the conclusions reached in § 5.

5. EVALUATION OF L_1 AND L_2

It has already been emphasized that D and E are not identical terms. When a polyatomic molecule is split into two parts by the rupture of a chemical link, the molecules or radicals formed reorganize themselves. Whether one or more of the atoms undergo a change in electron configuration with the entailed transference of energy or not, some change in the bond energies of the remaining links is to be expected, and this will also involve the absorption or liberation of energy. For the net effect we shall use the expression *energy of reorganization*, a term originally defined elsewhere (Norrish 1934). Only if the energy of reorganization is zero will D be equal to E.

As already mentioned, the experimental value of D(CH_3—H) at 25° C is 102 ± 1 kcal. (Andersen & Kistiakowsky 1943; Kistiakowsky & Van Artsdalen 1944). This implies that $D(CH_3—H)$ and $D(H_2)$ differ but slightly (about 2 kcal.), in good agreement with Patat (1936), who came to the conclusion that the reaction

$$CH_4+H = CH_3+H_2$$

is almost thermoneutral on the grounds that the forward and reverse processes require approximately equal activation energies.

The dissociation energy of the C_2H_5—H link is about 3 kcal. less than that of the CH_3—H link (Andersen & Van Artsdalen 1944). Since the actual C—H bond strength in the ethane molecule is only about 1 kcal. weaker than in CH_4 (table 3), this implies a difference of about 2 kcal. in the energies of reorganization of the ethyl and methyl radicals. No great quantitative significance should be attached to this figure as it is subject to relatively large experimental error. The value is reasonable, since it implies that $D(CH_3—CH_3)$ is 1·2 kcal. greater than $D(C_2H_5—C_2H_5)$, whereas a lessening of 1 kcal. in the reorganization energy difference would make $D(C_2H_5—C_2H_5)$ greater than $D(CH_3—CH_3)$, which scarcely seems likely in consideration of the fact that stability decreases rather than increases with increasing chain length. Since it might be expected that the energy of reorganization of the ethyl radical would be larger than that of the methyl radical, the observed difference infers that energy is liberated by these radicals at their formation from ethane and methane. Thus one would expect the energy required to remove the first hydrogen atom from CH_4 to be rather less than the average C—H bond energy, E(C—H) being slightly greater in CH_3 than in CH_4. It follows that the value of 102 kcal. for the energy of removing the first hydrogen atom from the methane molecule may be regarded as a minimum value for E(C—H), corresponding to a minimum figure of 182 kcal. for L_2. Since there is no change in the valence state of the carbon atom when one hydrogen atom is removed, and since the difference in electronegativity of carbon and hydrogen is small, no large energy of reorganization for the CH_3 radical is to be expected, and the difference in E(C—H) for CH_4 and CH_3 will be small. Our reasoning that the bond energy is the greater in CH_3 is supported by considerations brought forward by Van Vleck (1934). The actual bond energy in CH_4 cannot be fixed precisely, but it is not likely to differ from the weighted value of 104 kcal. by more than a small number of kilocalories in either direction. This figure leads to 190 kcal. for L_2. Since any variation in L_2 correspondingly affects E(C—H) by one-quarter of that amount, the uncertainty in L_2 is four times that in the C—H bond energy. The permissible limits for L_2 will be considered later.

If the estimate of Bacher & Goudsmit (1934) of 100 kcal. for the $^5S \rightarrow {}^3P$ conversion is correct, the value here suggested for L_2 implies a magnitude of about 90 kcal. for L_1. Since this would imply an equilibrium vapour pressure for carbon very much higher than has ever been experimentally indicated, it seems probable that Bacher & Goudsmit's figure is in excess of the true value. Reference has already been made to the alternative figures 125 and 170 kcal. for L_1, and it seems impossible to reconcile the higher of these with the foregoing observations; indeed, the value 170 kcal.

appears to be even less likely than a value lower than 125 kcal. The next lower value possible for correlation with the CO spectrum is 108·5 kcal., but even this value is very difficult to reconcile with the observed vapour pressure of carbon. On the other hand, a dimension of the order of 225 kcal. for L_2 would necessitate bond-energy values very much higher than those normally assigned to carbon compounds. The difference between L_2 and L_1 is therefore in all probability considerably less than 100 kcal., and $L_1 = 125·0$ kcal. at 0° K, corresponding to 9·144 eV for D_0(CO) (Herzberg 1939) and 27·20 kcal. for the heat of formation of CO at 0° K (Wagman, Kilpatrick, Taylor, Pitzer & Rossini 1945), still seems the most likely.

The employment of electron impact phenomena by Hagstrum & Tate (1941) to determine D(CO) has furnished the value 9·6 eV, necessitating about 136 kcal. for L_1. If the estimated appearance potential of the C^+ ion is in error it is in all probability high, and the result may be regarded as an upper limit. To reconcile the figure with that advocated by Herzberg from spectroscopic data presupposes an error of 0·5 eV, an amount exceeding that which Hagstrum & Tate are inclined to allow on the grounds that agreement with spectroscopic data is much closer in the case of other molecules: on the other hand, electron impact experiments have elsewhere indicated a somewhat lower value (8·8 eV) for D(CO) (Savard 1937). In any case, this evidence weighs decidedly against 170 kcal. as a possible value for L_1, and even more convincing is the photodecomposition of carbon monoxide by ultraviolet radiation of wave-length 1295 A with a quantum yield of unity (Faltings, Groth & Harteck 1938), which, as pointed out by Herzberg (1942), would appear to set a definite maximum of 135 kcal. on L_1.

Further confirmation that 125 kcal. is the most probable value for L_1 is to be found in the observations of Barrow (1940) concerning the linear relationship between D_0 and the internuclear distance for a series of diatomic molecules, each consisting of one atom in group IV and one in group VI (*b* subgroups) of the Periodic Table. If this relationship is valid for the whole series, $D_0(\mathrm{CO}) = 11·11$ eV, corresponding to $L_1 = 170$ kcal., would appear to be far too large to be permissible, 9·14 eV being the most satisfactory value. Undue emphasis is not placed on this point since the linearity is entirely empirical and only six of the values indicated were determined by rotational analysis, but attention is drawn to it as good circumstantial evidence.

The triple point of carbon has been determined by Basset (1939, 1942) who passed an electric current through a thin graphite rod under a varying pressure of argon until it either sublimed or melted. The co-ordinates of the triple point were found to be about 4000° K and 100 atm. Below this pressure the carbon sublimed without melting. Goldfinger & Jeunehomme (1936) have calculated the vapour pressures of carbon corresponding to each of the three values 107, 124 and 169 kcal. for L_1. For monatomic vapour the expected pressures at 4000° K are 151, 20 and 0·065 atm. respectively. The observed value is intermediate between the figures provided by the two lower values of L_1, agreement being best for $L_1 = 107$ kcal., but lies close to the total-pressure curve for $L_1 = 124$ kcal. for which the pressure of C_2 molecules has been taken into consideration. It is to be regretted that Basset does not provide an experimental sublimation curve. Presumably maximum temperatures, corresponding to the onset of sublimation, would have been observed at lower pressures, and his apparatus would have been suitable for carrying out the required measurements. Instead, he constructs a provisional phase diagram by combining his result with the readjusted figures obtained from measurements by other workers of the temperatures of the positive crater of the carbon arc at pressures of 1 atm. and below. The sublimation curve suggested cannot be correct, since, as can be demonstrated by the Clausius-Clapeyron relationship, it necessitates a heat of sublimation which increases rapidly with the temperature, reaching an impossibly high value long before the triple point is approached. The same author reports having obtained temperatures in the carbon arc considerably above the sublimation temperature at 22 kg./cm.^{2} pressure, and this supports previous criticisms of the arc method for determining the vapour pressure of carbon, on the grounds that the pressures indicated are considerably lower than the true equilibrium pressures.

Two of the rough measurements of Thiel & Ritter (1923) are derived by essentially equilibrium methods, and also support the intermediate value 125 kcal. for L_1. All the other experimental data referred to by Goldfinger & Jeunehomme (1936) either concern the dynamic method which, for reasons given in § 3, cannot be expected to provide direct information about L_1, or the arc method, in which true equilibrium conditions are not obtained.

The foregoing leads to the conclusion that the order of difference between L_2 and L_1 is about 65 kcal. and not 100 kcal. as estimated by Bacher & Goudsmit (1934). This is in accord with the theoretical calculations of Ufford (1938), who places the excitation energy of the ^5S carbon atom at about 73 kcal. It is also in accord with the conclusion reached by Burton, Davis, Gordon & Taylor (1941) concerning CH_2, namely, that if it has a molecular (unpromoted electron) structure, the heat of formation is only slightly above that of the radical (containing tetravalent carbon). The 65 kcal. liberated by the carbon atom in passing from the latter to the former structure is largely absorbed by the weakening of the two C—H bonds. Having taken E(C—H) for CH_4 to be 104 kcal. and that for CH_3 nearly 1 kcal. stronger, we would expect the two bonds in CH_2 to change from about 105 kcal. (or more) to about 80 kcal. (§ 3), thus absorbing approximately 50 kcal. In consequence, the energy required to raise the CH_2 molecule to the radical structure would be only some 15 kcal., in agreement with Burton *et al.*

Experimental evidence regarding the formyl radical also lends support to our interpretation. CHO, although not very stable, has been shown to exist at 100° C (Burton 1938). For the process

$$CHO \rightarrow H + CO,$$

Style & Summers (1946) calculate an activation energy of at least 16 kcal. Elsewhere the actual energy absorbed by this process is given as 26 kcal. (Gorin 1939). Whether these figures can be accepted uncritically or not, there is strong evidence that CHO possesses a definite stability which could scarcely be the case if $L_2 - L_1$ were as much as 100 kcal. The value of E(C—H) in CHO is not known with precision, but that it is very probably not more than 100 kcal. is indicated by the following facts. Atomic hydrogen reacts very readily with formaldehyde (Geib 1936) removing a hydrogen atom, but does not react with methane (Bonhoeffer & Harteck 1928), from which it appears that the dissociation energy of the first C—H bond has a smaller value in formaldehyde than in methane. Assuming that no considerable amount of energy is liberated by the reorganization of the CHO radical, it follows that the C—H bond strength in formaldehyde is likewise less than in methane. Since the force constants for the C—H bonds differ by some 10 % in the two compounds (Linnett 1945), being smaller in the case of formaldehyde, this may be regarded as a confirmatory indication that the C—H bond energy in CH_2O is appreciably weaker than in CH_4. In consequence our assumption that the reorganization of the CHO radical does not liberate more than a small amount of energy appears to be justified. Now the energy of dissociation of the first C—H bond in methane has been fixed at 102 ± 1 kcal. (Kistiakowsky & Van Artsdalen 1944). The foregoing evidence indicates that the corresponding dissociation energy in formaldehyde is somewhat smaller, and we regard it justifiable to assume that it is about 100 kcal. or rather less. Thus since the reorganization energy liberated by the CHO radical is small (or negative), it would also be expected that the values of E(C—H) in both CH_2O and CHO do not exceed 100 kcal.

When the hydrogen atom is removed from CHO, not only will energy be liberated by the switch of the carbon atom to the divalent state, but an additional amount released by the change in strength of the C=O linkage. This latter quantity is not small (Pauling 1939) and almost certainly exceeds 20 kcal. That the energy associated with this linkage is greater in CO than in CO_2 or the carbonyl group is an accepted fact, and further could be predicted from the shorter bond length, according to the observations of Wrinch & Harker (1940), who regard 'definite lengths as connoting definite energies'. In addition, Walsh (1946) has drawn attention to the existence of a smooth gradation of properties of the carbonyl bond in the series CO, CO_2, CH_2O, CH_3CHO and the ketones, even though the carbon in CO is divalent. The total variation is considerable. From the foregoing it is clear that if $L_2 - L_1 = 100$ kcal., the total energy available would exceed considerably that required to rupture the C—H bond, and CHO must needs be a very unstable and endothermic substance. But, on the contrary, the process

$$CO + H \rightarrow CHO$$

is very probably exothermic, since it is regarded as a necessary step in the formation of formaldehyde and glyoxal which occurs when mercury-sensitized mixtures of hydrogen and carbon monoxide are subjected to ultra-violet irradiation (Groth 1937; Faltings *et al.* 1938). The facts are very difficult to reconcile with a value higher than about 75 kcal. for $L_2 - L_1$. Unless L_1 has a value other than 125 kcal., this would place an upper limit of about 200 kcal. on L_2, in good agreement with our previous estimate of about 190 kcal. for L_2. The lower limit, as we have shown, is about 182 kcal. Thus to our figure of 190 kcal. for L_2 must be attached an uncertainty of some 10 kcal. in either direction. The corresponding value of $E(\text{C—C})$ is 95 ± 5 kcal. in diamond (§ 4) and 94 ± 5 kcal. in the normal paraffins (table 3).

Further light on the difference between L_2 and L_1 is forthcoming from the relative bond strengths in CO and CO_2. That the bond strength is definitely weaker in CO_2 is in accord with much experimental evidence, including the known facts regarding force constants, bond lengths and ionization potentials of the $p\pi$ electrons on the oxygen atom (Walsh 1946). Pauling (1939) explains this difference in terms of resonance energy, but his explanation cannot be accepted as it stands since he has not taken account of the energy difference in the valence states of carbon in the two compounds. Now the value we have accepted to be by far the most probable for $D_0(\text{CO})$ is 9·144 eV or 210·8 kcal. (Herzberg 1939). The heat of combustion of carbon monoxide, corrected to 0° K, is 66·77 kcal./mol. (Wagman *et al.* 1945). These figures can be combined with the value of $D_0(\text{O}_2)$ (Herzberg 1939) to calculate the mean bond energy in carbon dioxide at 0° K:

$$\begin{array}{rcll}
 & & & \text{kcal.} \\
\text{CO}+\tfrac{1}{2}\text{O}_2 &=& \text{CO}_2 & +\ 66{\cdot}77 \\
\text{C}(^3P)+\text{O} &=& \text{CO} & +210{\cdot}8 \\
\text{O} &=& \tfrac{1}{2}\text{O}_2 & +\ 58{\cdot}6 \\
\text{C}(^5S) &=& \text{C}(^3P) & +\ 65 \\
\hline
\text{C}(^5S)+2\text{O} &=& \text{CO}_2 & +401{\cdot}2
\end{array}$$

This is reasonable as it corresponds to an average value of 200·6 kcal. for $E(\text{C=O})$ in CO_2, that is, some 10 kcal. less than in CO. Had we taken $\text{C}(^5S) - \text{C}(^3P) = 100$ kcal., the bond energy in CO_2 would have assumed a mean value of 218·1 kcal., a figure exceeding that for CO. Thus, here again, experimental evidence is definitely against an estimate of 100 kcal. for the energy required by the $^3P \rightarrow {}^5S$ transformation, and indicates that the value 65 kcal. here suggested cannot be more than 10–15 kcal. below the true figure.

In like manner the strength of the carbonyl bond in formaldehyde can be estimated if a mean value is assumed for $E(\text{C—H})$ in this compound. The heat of formation of gaseous formaldehyde according to Delépine & Badoche (1942) is 25·7 kcal./mol. after a slight readjustment for the most recent data published on the heats of formation of CO_2 and H_2O has been effected. The slightly different value 27·7 kcal./mol. can be calculated from the figure for the heat of combustion at constant volume reported by von Wartenberg & Lerner-Steinberg (1925). The former claim greater accuracy, and the value adopted here as the 'best' in the absence of more precise data is 26·2 kcal./mol., which becomes 25·3 kcal./mol. at 0° K when allowance is made for the change in heat content. Combining this figure with other thermochemical quantities, the atomic heat of formation of formaldehyde at 0° K can be calculated as follows:

$$\begin{array}{rcll} & & & \text{kcal.} \\ \text{C (graphite)}+H_2+\tfrac{1}{2}O_2 &=& CH_2O\text{ (gas)} & +\ 25\cdot3 \\ O &=& \tfrac{1}{2}O_2 & +\ 58\cdot6 \\ 2H &=& H_2 & +103\cdot2 \\ C(^3P) &=& \text{C (graphite)} & +125\cdot0 \\ C(^5S) &=& C(^3P) & +\ 65 \\ \hline C(^5S)+2H+O &=& CH_2O\text{ (gas)} & +377\cdot1 \end{array}$$

Taking the value ~ 100 kcal. for the strength of the C—H bonds in formaldehyde, the implied strength of the carbonyl bond is ~ 177 kcal. Again, this is very reasonable, and far more so than the higher value provided by employing the old figure for the $^5S \rightarrow {}^3P$ transformation.

Finally, we wish to consider the work of Cherton* on acetylene which has come to our notice since we first deduced that L_2 is about 190 kcal. From a study of the products of decomposition, Cherton (1941) deduced that the C_2H_2 molecule initially loses a hydrogen atom in the electric discharge. Subsequently, it was shown that ultra-violet radiation has the same effect (Cherton 1942), and from the wave-length at which continuous absorption commences it was calculated that 121 kcal. are required to remove this hydrogen atom. This figure is reasonable, as here the C—H bond length is considerably shorter than for saturated hydrocarbons. Further, it agrees very well with the value supplied by an extrapolation of vibrational terms according to the method of Rydberg (Funke & Lindholm 1937), but, as might be expected, is considerably lower than that furnished by the Birge-Sponer extrapolation method. From this figure Cherton proceeds to calculate that the heat of sublimation of diamond to free atoms in the ^5S state is 190 kcal. At first sight this might appear to be excellent support for our value of 190 kcal. for L_2, but we are unable to proffer it as such as Cherton's calculation involves certain tacit assumptions. First, the atomic heat of formation of acetylene is calculated by adding the dissociation energies of the various linkages, the value taken for the C≡C bond being that derived from predissociation data (Price 1935). Quite apart from the doubt which has been cast in recent years upon the reliability of dissociation energy values obtained from predissociation in the spectra of polyatomic molecules, the treatment of dissociation energies as additive quantities is unsound. Only if the energies of dissociation are nearly equal to the bond strengths, that is, only if the energies of reorganization of the C_2H and CH radicals are small, will the result be a good approximation. Secondly, it is tacitly assumed that Price's evaluation of the energy required to rupture the C≡C linkage refers to a dissociation in which the carbon remains in the tetravalent state, that is, to the change

$$C_2H_2 = CH(^4\Sigma)+CH(^4\Sigma),$$

and not to

$$C_2H_2 = CH(^2\Pi)+CH(^2\Pi)$$

---

* We are indebted to M. Cherton for kindly providing us with reprints of his three papers, the original journals being unavailable in this country.

as supposed by Price. It is not impossible that Price's value does refer to the former change, as, allowing for the change in C—H bond strength, the energy level of the $^4\Sigma$ state will not exceed that of the $^2\Pi$ state by a prohibitively large amount. The energy required for the latter dissociation has been calculated to be 137·9 kcal. (Henkin & Burton 1940).

Thus Cherton's calculation is not rigid and does not provide closer limits for L_2. As we have shown, however, the figure he puts forward is of the right order, and furthermore, the bond energies calculated from it are compatible with experimental data on the pyrolysis of numerous hydrocarbons, as discussed in a third paper by the same author (Cherton 1943).

6. CONCLUSIONS

When the energy associated with the tetravalent state of carbon is systematically taken into account and a careful distinction made between bond energies and energies of dissociation, it is possible to bring order out of the relative state of chaos which has been associated hitherto with the thermochemistry of carbon. Practically all the experimental data from a wide variety of sources are now reconciled, and the value 190 ± about 10 kcal. for L_2, the latent heat of sublimation of carbon into free atoms at the lowest energy level of tetravalence (^5S), provides a scheme of bond energies for carbon compounds which is compatible with all the more important established quantities. This dimension for L_2, in combination with a considerable weight of other evidence, rules out the high value contested for L_1 (170 kcal.), a value supported from erroneous premises by those who fail to take into account the energy absorbed in the $^3P \rightarrow {}^5S$ conversion. There is very good evidence to indicate that this change involves considerably less than 100 kcal., the figure hitherto accepted, and by far the most probable and satisfactory value for L_1 is 125·0 kcal. at 0° K. The evidence may conveniently be summarized as follows:

evidence	deduction kcal.
Marshall & Norton's vapour-pressure measurements (dynamic method)	$L_2 \nless \sim 177$
Voge's calculations and $D(CH_3—H)$	$L_1 \leqslant 170$
Van Vleck's calculations and $D(CH_3—H)$	$L_2 > 182$
consideration of $D(CH_3—H)$, $D(C_2H_5—H)$ and energies of reorganization	$L_2 = \sim 190$, $L_1 \leqslant 170$
electron impact measurement of $D(CO)$	$L_1 \ngtr 136$
photodecomposition of CO	$L_1 \ngtr 135$
Barrow's empirical relationship	$L_1 = \sim 125$
triple point of graphite	$L_1 = \sim 125$
Thiel & Ritter's vapour-pressure measurements (equilibrium method)	$L_1 = \sim 130$
structure of CH_2	$L_2 - L_1 \leqslant 100$
stability of CHO	$L_2 - L_1 \ngtr 75$
relative $E(C=O)$ in CO, CO_2 and CH_2O	$L_2 - L_1 \ngtr 80$
pyrolysis data	$L_2 = \sim 190$
(spectroscopic determination of $D(CO)$	$L_1 = 125$)

The values of C—C and C—H bond energies for the various hydrocarbons, calculated as shown in table 3 assuming $L_2 = 190$ kcal., are, in our view, the most probable values, taking all evidence into account, and must be regarded as superseding bond energies calculated assuming other values for L_2, as, for example, in the recent paper of Skinner (1945).

APPENDIX

Considerations arising from the interpretation of the CN spectrum by Schmid, Gerö & Zemplén (1938) have suggested a more precise value for the quantity we have termed L_2. The CN dissociation limits, which are entirely experimental, having been obtained from short extrapolations, led these authors to the assignment of energy levels for the products of dissociation to agree with the belief that the excitation energy of the ^5S carbon atom is in the neighbourhood of 100 kcal.

In accord with our estimate (65 ± 10 kcal.) for the excitation energy of C(^5S), a reassignment of energy levels to the dissociation products of CN(A $^2\Pi$), namely C(^3P) + N(^2D) has indicated that the ^5S energy level for carbon is situated at about 24,200 cm.$^{-1}$ or 69·2 kcal. above the ground-level. Further details will be forthcoming in a later paper. These considerations are purely tentative, but, if correct, would provide the improved value $194{\cdot}2 \pm 4$ kcal. for L_2. In round figures this corresponds at 0° K to 104 ± 1 kcal. for the mean C—H bond energy in methane and 97 ± 2 kcal. for the single C—C link in solid carbon. The corresponding value for the dissociation energy of the C—C link in cyanogen is about 116 kcal., a very reasonable figure lying between the highly discordant extremes published in the literature. The acceptance of the foregoing brings the spectroscopic determination of D(CN) into alignment with Herzberg's determination of D(CO). The calculation of D(NC—CN) is based on the old accepted value for the dissociation energy of N_2 (Herzberg 1939). If, instead, we accept $D(N_2) = 225{\cdot}1$ kcal., the new value proposed by Gaydon & Penney (1945), the extremely high and unlikely value 171 kcal. results for D(NC—CN). We have already shown that the interpretation Gaydon & Penney assign to the CO spectrum is incorrect, and the indication is therefore that the same probably applies to their interpretation of the N_2 spectrum. Excellent confirmation of the forementioned value 116 kcal. for D(NC—CN) is forthcoming from the long wave-length limits of the second regions of continuous absorption at 1600 and 2100 A respectively in the spectra of CH_3CN (Herzberg & Scheibe 1930) and ICN (Badger & Woo 1931; Mooney & Reid 1931), these regions corresponding to photodissociation in which, as demonstrated by fluorescence, CN radicals are produced in the excited $B\,^2\Sigma$ state (Neuïmin & Terenin 1936; Yakovleva 1938).

REFERENCES

Andersen, H. C. & Kistiakowsky, G. B. 1943 *J. Chem. Phys.* **11**, 10.
Andersen, H. C., Kistiakowsky, G. B. & Van Artsdalen, E. R. 1942 *J. Chem. Phys.* **10**, 305.
Andersen, H. C. & Van Artsdalen, E. R. 1944 *J. Chem. Phys.* **12**, 479.
Bacher, R. F. & Goudsmit, S. 1934 *Phys. Rev.* ii, **46**, 948.

Badger, R. M. & Woo, S.-C. 1931 *J. Amer. Chem. Soc.* **53**, 2572.
Barrow, R. F. 1940 *Trans. Faraday Soc.* **36**, 1053.
Barrow, R. F., Pearson, T. G. & Purcell, R. H. 1939 *Trans. Faraday Soc.* **35**, 880.
Basset, J. 1939 *J. Phys. Radium*, vii, **10**, 217.
Basset, J. 1942 *Brennst-Chemie*, **23**, 127.
Baughan, E. C. 1941 *Nature*, **147**, 542.
Baughan, E. C. & Polanyi, M. 1940 *Nature*, **146**, 685.
Berriman, R. W. & Clark, C. H. D. 1938 *Proc. Leeds Phil. Lit. Soc.* (Sci. Sec.), **3**, 465.
Bonhoeffer, K. F. & Harteck, P. 1928 *Z. phys. Chem.* A, **139**, 64.
Brackett, F. S. 1928 *Proc. Nat. Acad. Sci., Wash.*, **14**, 857.
Burton, M. 1938 *J. Amer. Chem. Soc.* **60**, 212.
Burton, M., Davis, T. W., Gordon, A. & Taylor, H. A. 1941 *J. Amer. Chem. Soc.* **63**, 1956.
Cherton, R. 1941 *Bull. Soc. Sci. Liége*, **10**, 604.
Cherton, R. 1942 *Bull. Soc. Sci. Liége*, **11**, 203.
Cherton, R. 1943 *Bull. Soc. Chim. Belg.* **52**, 26.
Conn, J. B., Kistiakowsky, G. B. & Smith, E. A. 1939 *J. Amer. Chem. Soc.* **61**, 1868.
Cook, M. A. 1945 *J. Chem. Phys.* **13**, 262.
Delépine, M. & Badoche, M. 1942 *C.R. Acad. Sci., Paris*, **214**, 777.
Eyring, H. 1932 *J. Amer. Chem. Soc.* **54**, 3191.
Faltings, K., Groth, W. & Harteck, P. 1938 *Z. phys. Chem.* B, **41**, 15.
Funke, G. W. & Lindholm, E. 1937 *Z. Phys.* **106**, 518.
Gaydon, A. G. & Penney, W. G. 1945 *Proc. Roy. Soc.* A, **183**, 374.
Geib, K. H. 1936 *Ergebn. exakt. Naturw.* **15**, 44.
Giauque, W. F. 1930 *J. Amer. Chem. Soc.* **52**, 4816.
Goldfinger, P. & Jeunehomme, W. 1936 *Trans. Faraday Soc.* **32**, 1591.
Gorin, E. 1939 *J. Chem. Phys.* **7**, 256.
Groth, W. 1937 *Z. phys. Chem.* B, **37**. 315.
Hagstrum, H. D. & Tate, J. T. 1941 *Phys. Rev.* ii, **59**, 354.
Henkin, H. & Burton, M. 1940 *J. Chem. Phys.* **8**, 297.
Herzberg, G. 1937 *Chem. Rev.* **20**, 145.
Herzberg, G. 1939 *Molecular spectra and molecular structure.* I. *Diatomic molecules.* New York: Prentice-Hall.
Herzberg, G. 1942 *J. Chem. Phys.* **10**, 306.
Herzberg, G., Herzfeld, K. F. & Teller, E. 1937 *J. Phys. Chem.* **41**, 325.
Herzberg, G. & Scheibe, G. 1930 *Z. phys. Chem.* B, **7**, 390.
Holcomb, D. E. & Brown, G. G. 1942 *Industr. Engng Chem.* **34**, 590.
James, H. M. & Coolidge, A. S. 1933 *J. Chem. Phys.* **1**, 825.
Jessup, R. S. 1937 *Bur. Stand. J. Res., Wash.*, **18**, 115.
Kistiakowsky, G. B. & Van Artsdalen, E. R. 1944 *J. Chem. Phys.* **12**, 469.
Kynch, G. J. & Penney, W. G. 1941 *Proc. Roy. Soc.* A, **179**, 214.
Lasareff, W. 1935 *Physica*, New Ser. **2**, 737.
Linnett, J. W. 1945 *Trans. Faraday Soc.* **41**, 223.
Marshall, A. L. & Norton, F. J. 1933 *J. Amer. Chem. Soc.* **55**, 431.
Mecke, R. 1930 *Z. Elektrochem.* **36**, 589.
Mooney, R. B. & Reid, H. G. 1931 *Nature*, **128**, 271.
Mulliken, R. S. & Rieke, C. A. 1942 *Rev. Mod. Phys.* **14**, 259.
Neuîmin, H. H. & Terenin, A. N. 1936 *Acta physicochim. U.R.S.S.* **5**, 465.
Norrish, R. G. W. 1934 *Trans. Faraday Soc.* **30**, 103.
Patat, F. 1936 *Z. phys. Chem.* B, **32**, 274.
Pauling, L. 1939 *The nature of the chemical bond.* Cornell Univ. Press.
Pearson, T. G., Purcell, R. H. & Saigh, G. S. 1938 *J. Chem. Soc.* p. 409.
Pitzer, K. S. 1940 *Chem. Rev.* **27**, 39.
Price, W. C. 1935 *Phys. Rev.* ii, **47**, 444.
Prosen, E. J., Jessup, R. S. & Rossini, F. D. 1944 *Bur. Stand. J. Res., Wash.*, **33**, 447.
Prosen, E. J., Pitzer, K. S. & Rossini, F. D. 1945 *Bur. Stand. J. Res., Wash.*, **34**, 403.
Prosen, E. J. & Rossini, F. D. 1945 *Bur. Stand. J. Res., Wash.*, **34**, 263.

Rossini, F. D. 1931 *Bur. Stand. J. Res., Wash.*, **7**, 329.
Rossini, F. D. 1934*a* *Bur. Stand. J. Res., Wash.*, **12**, 735.
Rossini, F. D. 1934*b* *Bur. Stand. J. Res., Wash.*, **13**, 21.
Rossini, F. D. 1940 *Chem. Rev.* **27**, 1.
Savard, J. 1937 *Rev. faculté sci. univ. Istanbul,* New Ser. **2**, 235.
Schmid, R. 1936 *Z. Phys.* **99**, 274.
Schmid, R., Gerö, L. & Zemplén, J. 1938 *Proc. Phys. Soc.* **50**, 283.
Shidei, T. 1936 *Jap. J. Phys.* **11**, 23.
Sidgwick, N. V. & Bowen, E. J. 1931 *Ann. Rep. Chem. Soc.* p. 367.
Skinner, H. A. 1945 *Trans. Faraday Soc.* **41**, 645.
Smith, J. O. (Jnr) & Taylor, H. S. 1939 *J. Chem. Phys.* **7**, 390.
Stevenson, D. P. 1942 *J. Chem. Phys.* **10**, 291.
Style, D. W. G. & Summers, D. 1946 *Trans. Faraday Soc.* **42**, 388.
Sutherland, G. B. B. M. 1938 *Proc. Indian Acad. Sci.* A, **8**, 341.
Sutherland, G. B. B. M. 1940 *J. Chem. Phys.* **8**, 161.
Thiel, A. & Ritter, F. 1923 *Z. anorg. Chem.* **132**, 125.
Ufford, C. W. 1938 *Phys. Rev.* ii, **53**, 568.
Van Artsdalen, E. R. 1942 *J. Chem. Phys.* **10**, 653.
Van Vleck, J. H. 1934 *J. Chem. Phys.* **2**, 20.
Voge, H. H. 1936 *J. Chem. Phys.* **4**, 581.
von Wartenberg, H. & Lerner-Steinberg, B. 1925 *Z. angew. Chem.* **38**, 591.
Wagman, D. D., Kilpatrick, J. E., Taylor, W. J., Pitzer, K. S. & Rossini, F. D. 1945 *Bur. Stand. J. Res., Wash.*, **34**, 143.
Walsh, A. D. 1946 *Trans. Faraday Soc.* **42**, 56.
Wrinch, D. & Harker, D. 1940 *J. Chem. Phys.* **8**, 502.
Yakovleva, A. V. 1938 *Acta physicochim. U.R.S.S.* **9**, 665.

The experimental study of the blast from bombs and bare charges

By G. Grime, M.Sc. and H. Sheard, M.Sc.

(*Communicated by Sir Edward Appleton, F.R.S.—Received* 3 *October* 1944)

The paper gives an account of investigations on the blast from bare and cased charges of explosive ranging in weight from a few pounds to several thousand pounds.

The measuring technique, involving the use of piezo-electric gauges with cathode-ray oscillographs, is described, and features of theoretical interest or of practical importance are illustrated.

It is shown that, at any rate at distances large compared with the dimensions of the charge, the scale relationship, deduced from simple dimensional theory, holds good over a wide range of charge weight. The importance of this finding in relation to model and full-scale investigations of bomb behaviour is discussed.

Comparisons of the measured velocity of the blast wave with that calculated, by the Rankine-Hugoniot formula, from the maximum excess pressure behind the shock-front show that the recorded values are in good agreement with those deduced theoretically. Observed and calculated rates of decay of maximum excess pressure are also in reasonable agreement.

INTRODUCTION

Just prior to the outbreak of war in 1939 the Home Office (A.R.P. Dept.) sought information from the Building Research Station of the Department of Scientific and Industrial Research as to the effect on buildings of the blast from bombs. A mathematical analysis, carried out by Dr E. N. Fox, showed that certain basic data were required. A search of the available literature yielded very little of value; the results of a few measurements of the blast pressure from explosive charges had been published (see ref. (1)), but most of the measurements had been made so far away from the charge that no use could be made of them in an analysis concerned with damage. To produce the required information, a research was therefore initiated, first, to devise accurate and convenient methods of recording the pressure-time curves near explosions, and, secondly, to use these methods to obtain blast data for a representative series of bombs and charges.

The apparatus, developed as the result of work started at the Building Research Station, and continued at the Road Research Laboratory of the Department of Scientific and Industrial Research after the outbreak of war, has been used to measure the blast pressures from explosive charges ranging from a few ounces in weight to the largest bombs in Service use. Similar apparatus now provides a standard method of assessing the blast performance of new munitions.

The present paper, having described the apparatus, reviews some of the results obtained, up to the end of December 1942, in the course of investigations undertaken by the Road Research Laboratory, initially for the Research and Experiments Department of the Ministry of Home Security, and later for the Ministry of Supply.

EXPERIMENTAL

At the outset it was known that the initial pressure rise in a blast wave is very steep, and it was realized that mechanical methods of recording were likely to be unsatisfactory. The following requirements have to be met by a satisfactory recording apparatus:

(1) the sensitive element or gauge must have a high natural frequency, since the initial pressure rise in the blast wave is very steep;

(2) the gauge must be capable of use with cables at least 100 yd. long;

(3) the recording apparatus must deal with several records simultaneously; and

(4) it must record without distortion the highest frequency accurately reproduced by the gauge and amplifier;

(5) it is essential that the whole apparatus be as simple as possible, since most of the work is done under field conditions.

Of several electrical methods of pressure recording considered, one depending on the piezo-electric properties of quartz was finally adopted, and the apparatus was designed to satisfy the above requirements. The piezo-electric quartz gauge, developed for the work, has a natural frequency of about 40,000 (type 1) or 70,000 (type 2) cyc./sec., and can be used with leads 500 yd. long; the recording apparatus, which is of very simple design, requires no external power supplies, and each recording unit gives photographic registration of the output from three gauges on separate cathode-ray oscillographs. Several such recording units may be operated simultaneously.

The gauges. An appropriately cut piezo-electric crystal, when compressed or elongated, develops a charge on two of its faces; the quantity of charge is proportional to the applied load and is developed without appreciable time delay. To utilize the phenomenon, the charge, collected on electrodes cemented to the two faces, is applied to a condenser of known capacity and the resulting voltage is recorded. Quartz or tourmaline is generally used for pressure recording, and the gauges described below are designed for quartz crystals.

Figure 1 is a drawing of the simplest form of blast pressure gauge (type 1), while figure 2 shows an improved type (type 2), with double the sensitivity of the first. Photographs of the two gauges are reproduced in figure 3.

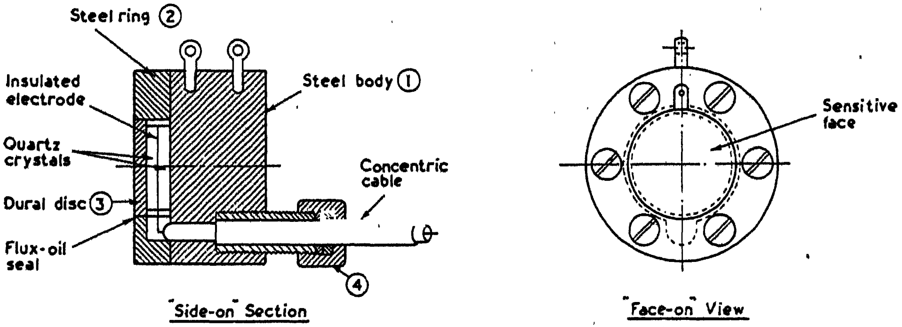


FIGURE 1. Piezo-electric blast pressure gauge—type 1.

The main body (1) of the type 1 gauge is a cylindrical block of steel, 2 in. in diameter by 1 in. thick, through which a single-core shielded cable (Telcon K.I.C.) is brought and secured by a fitting (4). Two circular (X-cut) quartz plates are mounted on the front plane surface of (1) and are covered by a duralumin plate (3) of slightly greater diameter. Between the crystals is placed an electrode of copper foil, to which the central wire of the cable is soldered. The pile of disks, consisting of the duralumin plate, the two crystals, and the central electrode are all cemented together and to the block (1) with a high melting-point bitumen. A steel ring (2), screwed to the block (1), fits round the plate (3), the annular space between the plate and ring (about

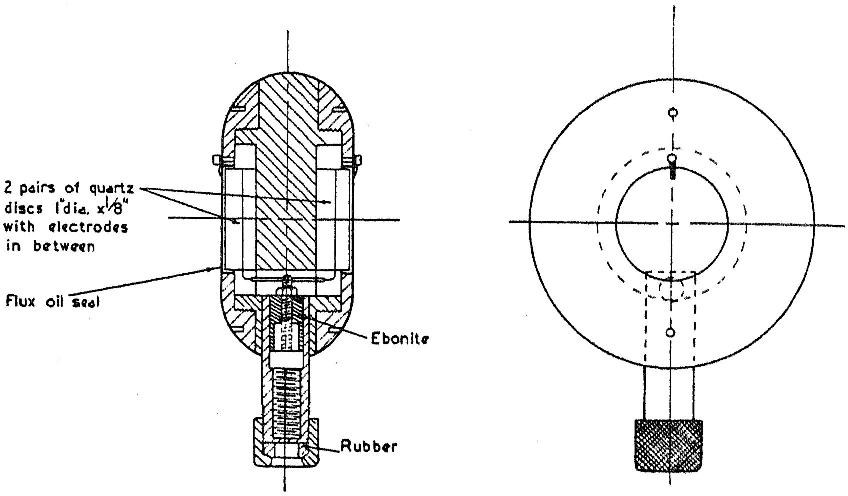
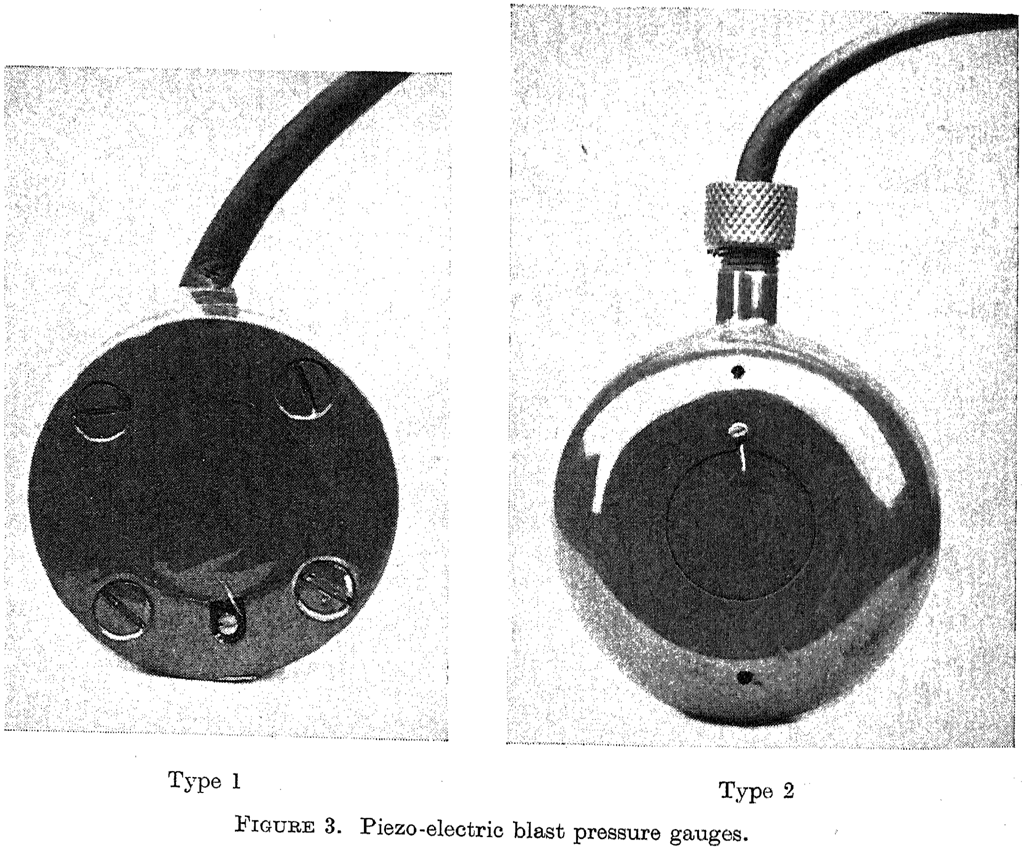

2 pairs of quartz discs 1"dia. x 1/8" with electrodes in between
Flux oil seal
Ebonite
Rubber

FIGURE 2. Blast pressure gauge—type 2.



Type 1

Type 2

FIGURE 3. Piezo-electric blast pressure gauges.

1/1000 in. wide) being filled with a very viscous oil, which serves the double purpose of forming an airtight seal and of providing damping for the crystals.

The lowest natural frequency of 40,000 cyc./sec. from a gauge of this type is very probably due to flexural vibration of the steel block or ring, since the frequency of the longitudinal front-to-back vibration is found by calculation to be more than twice this figure.

Type 2 is similar in essentials but has two pairs of crystals, one on either side of a central steel plate. The crystals are connected electrically in parallel, so that for pressures the charges reinforce one another; as regards vibration of the central plate, however, the two pairs act in opposition, and no net charge results. The natural frequency of the type 2 gauge is about 70,000 cyc./sec.

Both types of gauge are used with their sensitive faces parallel with the direction of propagation of the blast.

When bare gauges are relatively near to a charge there is some transmission of heat from the blast wave to the sensitive face of the gauge. This tends to cause distortion of the crystal pile, giving rise to parasitic effects. A layer of vaseline over the face of the gauge hinders this heating and eliminates the trouble.

The recording unit. The complete recording unit, including three amplifiers and oscillographs, a drum camera and all battery supplies, is contained in two wooden boxes measuring respectively 35 × 22 × 15 in. and 31 × 18 × 6 in. (figure 4).

The arrangement of the electrical circuit of one of these triple units is given in figure 5.

The two-stage resistance capacity coupled amplifiers employing high-frequency pentodes (Mullard S.P.2's) give a voltage amplification of about 8000. The good low-frequency response, necessary to record accurately the long-duration blast waves from large bombs, is obtained by separating the anode and screen supplies to the first and second stages; all three first stages are fed from one battery and the second stages from another. The high-frequency response is adequate in relation to the characteristics of the gauge, the loss at 10,000 cyc./sec. being about 4 %. One other feature of the amplifiers calls for comment; in the first stage of each amplifier, it is necessary to have a valve of high grid-filament resistance, to reduce leakage from the piezo-crystal to a minimum. Valves are considerably improved in this respect by under-running their filaments, and a good high-frequency pentode with a 3-ohm resistance in one valve leg, as shown in the circuit diagram, is very satisfactory, and will operate in a stable manner with a grid leak as high as 200 megohms. The arrangement has the further advantage of providing grid bias for the first stage. A 2 V accumulator supplies filament current for the three amplifiers, and three 120 V dry batteries

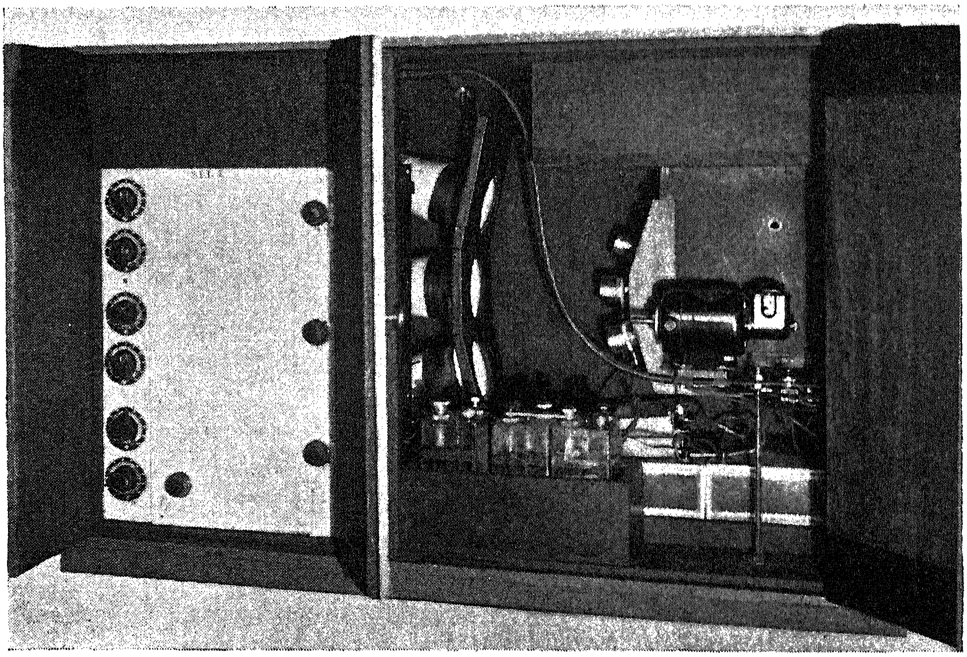


FIGURE 4*a*. Triple oscillograph unit.

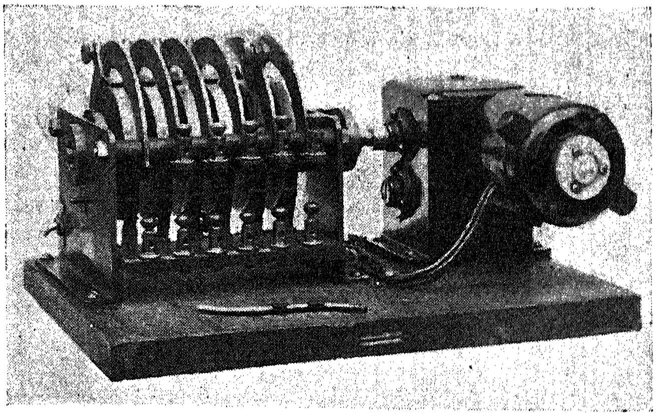
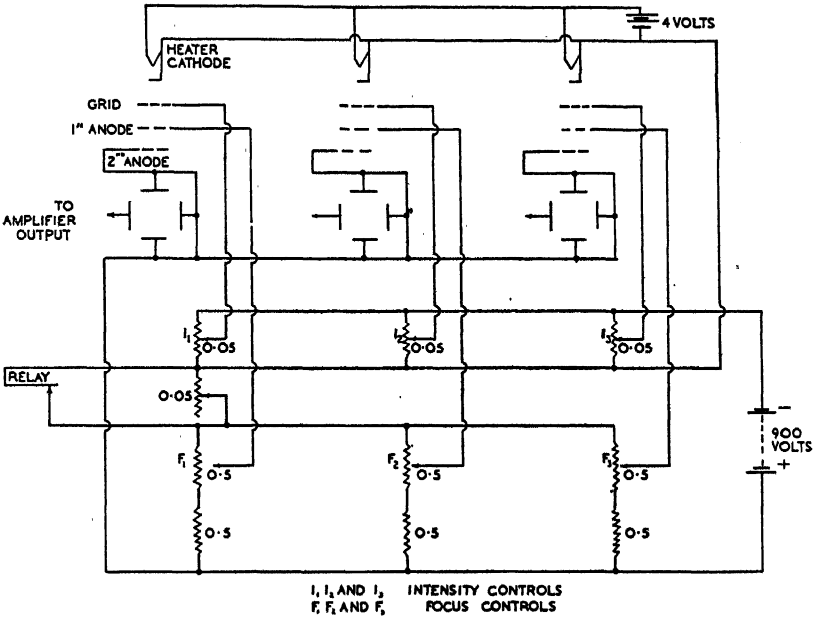


FIGURE 4*b*. Rotary switch.

supply anode current, one for the first stages, and two for the second stages. The amplifiers provide a substantially linear output of about 100 V.

The oscillographs at present in use (G.E.C. V.C.R. 138's) have $3\frac{1}{4}$ in. photographic screens. A high-tension voltage of 900 V is provided from dry



CATHODE RAY TUBE CIRCUIT

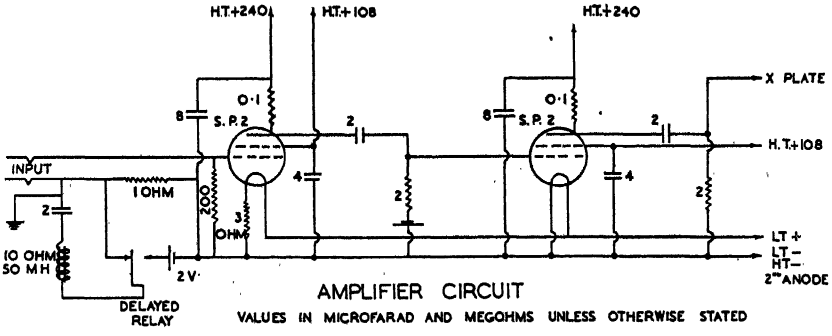


FIGURE 5. Electrical circuit of a triple oscillograph unit.

batteries housed in a separate wooden box, which also contains a 100 V battery for the operation of relays. The fluorescent spots of the three oscillographs are normally suppressed by the negative bias on the intensity control electrode (figure 5) until the moment of recording, but, during the recording period, they are brought up to full intensity by a relay, which short-circuits the resistor which provides this negative bias. The relay coil has a resistance of about 1000 ohms and may be operated from any convenient distance, since the resistance of any reasonable length of connecting cable is small compared with that of the relay coil.

The recording spots are deflected in a horizontal direction only, and a time axis is provided by registering all three records on a 4 in. wide strip of photographic film or paper fastened round a 7 in. diameter drum in a removable camera. The camera is fitted with a dark slide, which, when open, allows the spots to be focused on the film by Kodascope F 1·6 lenses of 2 in. focal length. The drum is driven by a 6 V Klaxon motor, geared down to a speed of 2 rev./sec.

Synchronization of the recording and firing is accomplished by a motor-driven switch with rotating contacts (figure 4), which first operates the brightness control relay and then fires the explosive charge.

Calibration. Two methods have been used for the calibration of gauges. In the first air pressure is suddenly applied to the gauge. The apparatus (figure 6) consists essentially of two chambers separated by a thin diaphragm of cellophane or similar material. The gauge, coated with vaseline, is clamped over a hole in the bottom of the lower chamber (type 2 is placed inside the chamber), and is connected to a standard condenser and to the oscillograph in the usual manner. The upper chamber is pumped up by means

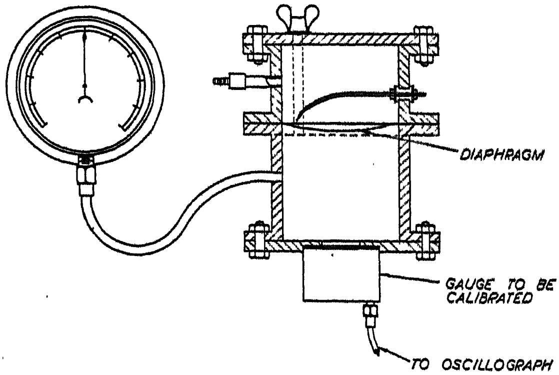
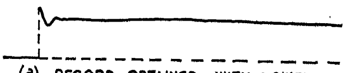
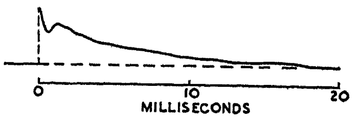


FIGURE 6. Diagram of calibration apparatus.



(a) RECORD OBTAINED WITH LOWER CHAMBER OF CALIBRATION APPARATUS PACKED WITH RAG



(b) RECORD OBTAINED WHEN INSULATION RESISTANCE OF GAUGE IS FAULTY

FIGURE 7. Typical calibration records.

of a bicycle pump, and the diaphragm is then punctured by a sharp point operated from outside. The diaphragm shatters and the pressure is applied suddenly to the gauge, the value of the final steady pressure being read off on a pressure gauge of the Bourdon type connected to the lower chamber. In order to ensure that the pressure change on expansion is substantially isothermal, and to damp out the natural frequency of vibration of the air in the apparatus, the lower chamber is packed fairly tightly with rag, when a trace of the form shown in figure 7 is produced. The vertical distance between the initial zero line and the deflected line is the deflexion due to the final steady pressure read from the Bourdon gauge. This gauge has been calibrated against a mercury column, and found to be accurate to within 1/10 lb./sq.in. at all values of pressure.

In addition to providing a gauge calibration, the method has the advantage that the slope of the oscillograph trace after the application of the steady pressure gives information about the time constant of the whole assembly—gauge and amplifier.

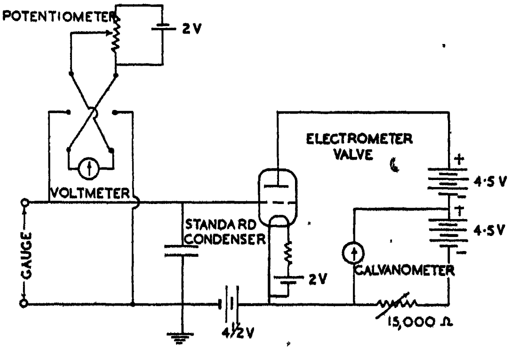


FIGURE 8. Electrical circuit of valve voltmeter used in static calibration.

Instead of making use of the recording equipment for the calibration, it is generally more convenient and rather more accurate to determine the sensitivity of the gauge alone, that of the recording equipment being registered automatically in the manner described in the next paragraph. The gauge to be calibrated is subjected to a steady pressure, either by placing known weights on the sensitive face, or by applying known air pressures with the apparatus illustrated in figure 6; the latter method is preferable, since, for highest accuracy, it is essential to apply the calibrating pressure uniformly. The resulting voltages, developed across a standard condenser in parallel with the gauge, are measured by means of a valve voltmeter; a suitable circuit is shown in figure 8. For successful operation, the grid-filament resistance of the valve employed, and the insulation resistance of the gauge should be very high, so that errors due to the leakage of charge from the gauge during calibration are negligible.

Each time a record is made, the amplifier and oscillograph sensitivity are checked by an automatic calibrating circuit, which, immediately after the oscillograph spots are brought to full brightness, injects a damped sinusoidal voltage of known value into the earth lead. The calibrator consists of a choke, a condenser, and a 2 V cell, all in series with a resistor of a few ohms inserted in the earth lead common to all the amplifiers in use. It is actuated by a delayed relay in parallel with that bringing the spots to recording brightness.

In addition to its function of defining the sensitivity of the amplifier and oscillograph, the calibrating oscillation at the commencement of each record furnishes a datum point for the measurement of time, and, since the frequency of oscillation is known, it also enables the time scale to be calculated.

Accuracy of measurement. Factors influencing the accuracy of measurements of maximum pressure are:

(1) *The natural frequency of the gauge.* Calculations of the amplitude of the oscillations at the natural frequency of the gauge are rendered difficult by the fact that the lowest natural frequency appears to be due, not to a thickness vibration of the crystal pile, but to a flexural vibration of the steel housing, excited by the impact of the blast on the gauge. In circumstances favourable to its excitation, however, the characteristic frequency may be detected on the record. To reduce errors caused thereby, a thin layer of sorbo rubber is wrapped round the curved surface of the gauge, leaving the faces carrying the sensitive disks exposed, when the amplitude is reduced to about 1 % of the maximum deflexion and is too small to be measured accurately. (This precaution is not always observed in bomb trials, where, as shown later, the maximum pressures may be relatively unimportant. Figures 9 and 10, for example, were obtained without sorbo screens on the gauges.)

(2) *The diameter of the sensitive disk.* The normal method of setting up a blast pressure gauge being to suspend it with the sensitive face or faces parallel with the direction of travel of the blast wave, it is clear that the larger the diameter of the sensitive disk, the poorer is the reproduction of the fine structure of the blast wave. The extension of the gauge along the direction of travel of the blast results in a reduction of the slopes of lines and a rounding off of peaks in the blast record. In particular, the recorded maximum pressure, assuming no error due to gauge frequency, is always less than the true peak pressure.

The magnitude of the correction to the peak pressure to take account of the disk size has been estimated by assuming, for the positive phase of the blast wave,* a shock front followed by an exponential decay curve.

Then, if the decay curve is represented by

$$p = p_1 e^{-nt},$$

---

* See next section.

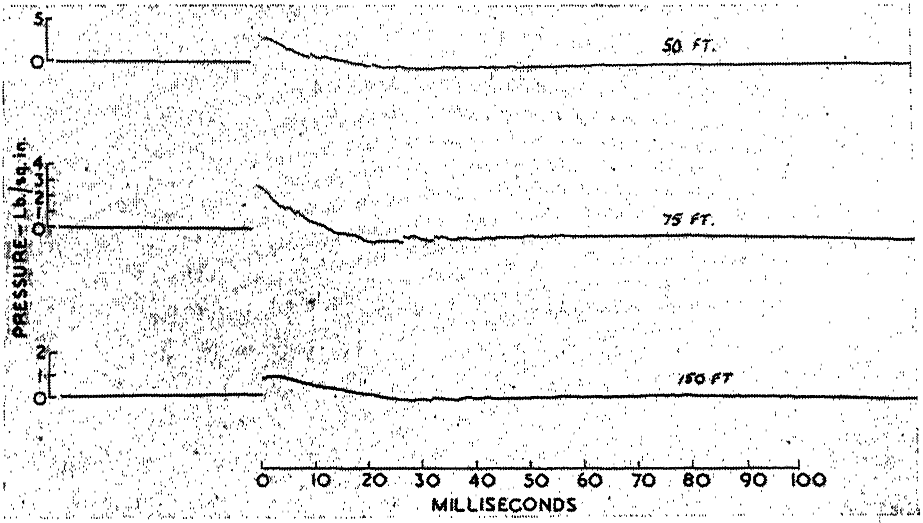


FIGURE 9. Blast pressure records from a bare charge of 70 lb. 60/40 R.D.X./T.N.T.

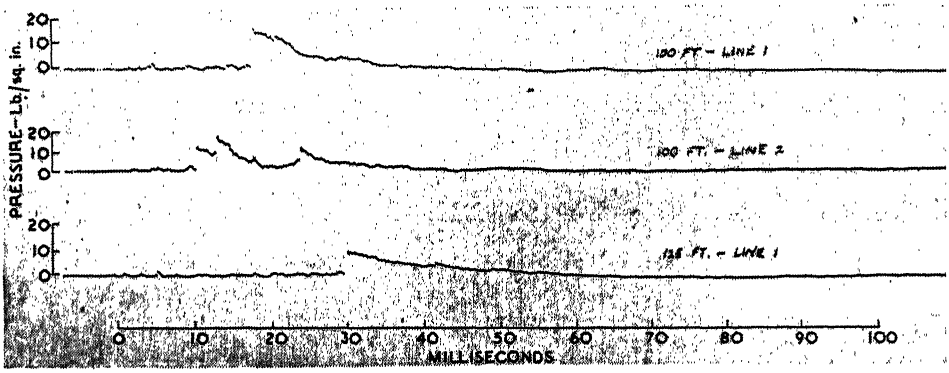


FIGURE 10. Blast pressure records from a British 4000-lb. H.C. bomb filled 60/40 amatol—100 and 125 ft.

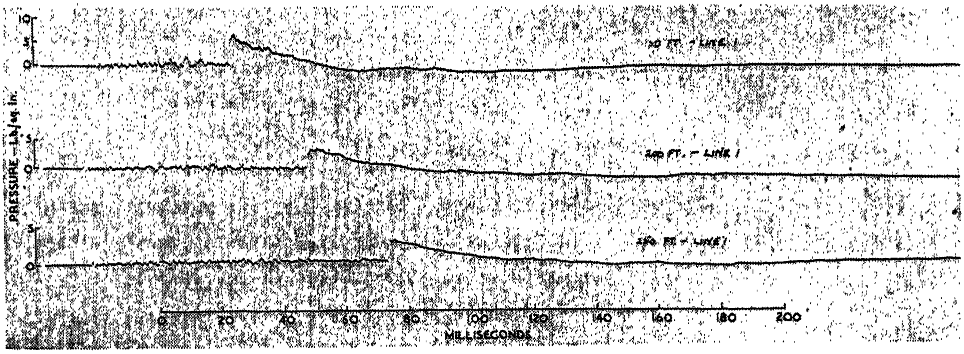


FIGURE 10 (*cont.*). Blast pressure records from a British 4000-lb. H.C. bomb filled 60/40 amatol—150 to 250 ft.

where p is the pressure at time t after the arrival of the shock front and p_1 is the maximum pressure in the wave, the recorded maximum pressure

$$p_r = p_1\left(1-\frac{an}{c}\right) \text{ approximately,}$$

where a is the radius of the sensitive disk, and c the velocity of the shock wave. The approximate correction given above is sufficiently accurate for the small corrections which are required in practice; as might be anticipated, the correction is mainly determined by the time taken for the shock front to pass over the disk in relation to the initial slope of the decay curve.

(3) *The frequency response of the amplifier.* Lack of response at high frequencies in the amplifier gives rise to a rounding off of peaks in the pressure record. An analysis, by A. R. Bryant (see ref. (2)), enables an estimate to be made of the reduction of the maximum value of the signal from the gauge in its passage through the amplifier.

Starting, as before, with a pressure wave of the form $p = p_1 e^{-nt}$, he shows that the peak output voltage

$$v_m = v_g \epsilon (k)^{-n\tau/(1-n\tau)},$$

where v_g = peak input voltage from the gauge,

ϵ = amplification factor at the middle of the frequency range,

$\tau = \dfrac{1}{2\pi f_0}$ for a single-stage amplifier,

$= \dfrac{1}{9{\cdot}75 f_0}$ for a two-stage amplifier with similar stages,

f_0 = upper frequency at which the amplification is $0{\cdot}707\epsilon$,

$k = 1+\dfrac{1-n\tau}{n\tau}\,\dfrac{c\tau}{2a}\left(1-e^{-2a/c\tau}\right).$

For small corrections, $n\tau$ and $e^{-2a/c\tau}$ are negligible compared with unity, and the expressions may be simplified to

$$v_m = v_g \epsilon\left[1+\frac{1}{2}\frac{c}{an}\right]^{-n\tau}.$$

The combined correction for gauge size and frequency response is therefore

$$p_r = p_1\left(1-\frac{an}{c}\right)\left(1+\frac{1}{2}\frac{c}{an}\right)^{-n\tau}.$$

Unless very sharp peaks occur in the record, the corrections are of importance only for small charges; for example, in a particular test with a 500 lb. M.C. bomb, at a distance of 40 ft., it was found that $n = 1100$ and $c = 1800$ ft./sec. a was 0·5 in., and f_0 was 30,000 cyc./sec., so that

$$\begin{aligned} p_r &= p_1 \times 0{\cdot}974 \times 0{\cdot}989 \\ &= 0{\cdot}965 p_1. \end{aligned}$$

(4) *Transmission-line phenomena in the cable.* Mention must also be made of transmission-line phenomena, as, at first sight, they might appear to be of some importance in tests of large bombs, where it is necessary to use long cables to connect the gauges to the amplifiers. The usual transmission-line formulae are not directly applicable to the problem, as they assume a constant voltage generator of sinusoidal wave-form, whereas the piezo-electric gauge produces a transient charge. Calculations, based on the cable constants, of the velocity of a transient along the cable, however, indicate that line phenomena are unlikely to be of major importance even when long cables are in use. With the particular cable employed, this velocity is found to be about 490×10^6 ft./sec. The time taken to traverse the longest cable in use, 500 yd. in length, is therefore about $3\,\mu$sec. Assuming similar reflecting conditions at each end of the cable, the natural frequency of a 500 yd. cable is therefore about 166,000 cyc./sec., and the effect of the cable is to introduce into the amplifier disturbances recurring at this frequency. As the time of rise to maximum voltage input to the amplifier is about 1/12,000 sec., that is, about 14 times that taken for the wave to travel to and fro in the cable, the elimination of these line oscillations by attenuation in the amplifier is unlikely to be attended by any serious loss of accuracy.

The factors dealt with in the preceding paragraphs lead to systematic errors for which corrections have to be made. The various operations of calibration and recording also give rise to random errors, the magnitudes of which have been examined by carrying out these operations many times. The standard error of a single measurement of maximum pressure, determined in this way, was found to be about 2·8 %.

With regard to the measurement of positive impulse, that is, the time integral of the positive phase, unless special precautions are taken, by far the most serious source of error is likely to be the generation of parasitic charges by violent movements of the cable connecting the gauge to the amplifier. Much can be done to reduce this by the correct choice of cable, and, in particular, by avoiding cables insulated with plastics; with the cable used in the present work, the errors due to cable signal are negligible except at positions very near to large bombs, where the gauges may be displaced several feet by the blast.

When pressure waves of considerable duration are recorded, it is desirable to arrange that the time constants of the gauge and amplifier circuits shall be long. It may readily be shown that if the positive phase is triangular, with a duration t_1, and the gauge has a parallel capacity C and resistance R, i.e. a time constant CR, the ratio of the recorded impulse I_r to the impulse I delivered by a gauge with infinite time constant is

$$\frac{I_r}{I} = 1 - \frac{2}{3}\frac{t_1}{CR} + \frac{1}{2}\frac{t_1^2}{C^2R^2} \text{ approximately,}$$

assuming no distortion in the amplifier. A similar expression describes the behaviour of each stage of the amplifier, C in this case being the interstage coupling capacity and R the sum of the anode and grid resistances.

The time constants of the gauge circuit and of each stage of the amplifier are each about 4 sec. For a 500 lb. M.C. bomb, where the duration of the positive phase may be about 15 msec., the ratio $I_r/I = 0{\cdot}9975$ for the gauge circuit. The overall error in the impulse measurement, including the contribution of the amplifier, if no correction were applied, would therefore be about 0·75 %.

A closer determination of the experimental error may be made for impulse than for maximum pressure, and, assuming negligible cable signal, the standard error of a single observation of the positive impulse from a bomb has been found to be ± 3 %.

RESULTS

Measurements have been made of the blast pressures produced by bare charges up to 2000 lb. in weight, by German bombs and mines weighing from 50 to 1000 kg., and by British bombs of all sizes.

Typical records. The normal type of pressure-time curve from a bare charge at all distances at which records have been obtained, and at considerable distances from a cased charge or bomb, consists of a shock-fronted positive phase, roughly triangular in shape, followed by a suction phase having a duration several times that of the positive and a lower maximum. The suction phase usually approximates in shape to one-half of a damped sine wave, but frequently departs from this normal form. Figure 9 illustrates pressure-time curves of this type.

Near a cased charge or bomb it is unusual to record a positive phase of this simple form; instead, the positive may consist of two or more shock-fronted waves superimposed (see figure 10), or may even be made up of a series of short-duration shock-fronted waves superimposed on a rounded pressure pulse. As the wave travels outwards, however, it gradually assumes a shape similar to that characteristic of a bare charge.

All records of the pressure from cased charges exhibit, in addition to the main features discussed above, numbers of small bow waves due to the fragments of the case. At all distances where measurements have hitherto been made, the fragments are ahead of the blast, so that the fragment bow waves are always recorded before as well as with the main blast wave.

The scale relationship. From dimensional theory, it can be shown that, for geometrically similar charges of the same composition and density, the blast pressure-time curves from different charges should appear identical if the units of distance and time adopted are proportional to the linear dimensions of the charge. It follows that, if the maximum value of the excess pressure is plotted against the distance from the charge, taking the charge diameter as the unit of distance, the results for similar charges of different weights should all lie on one curve. The durations of the positive and negative phases at corresponding distances, i.e. at the same number of diameters from the charge, should be proportional to the respective charge diameters. Thus if the durations, divided by the appropriate charge diameters, are plotted against distance in charge diameters, the points should again fall on one curve. The impulse, i.e. the area under the pressure-time curve, clearly scales in a manner similar to the duration.

The scale relationship has been found to hold well, at any rate at distances large compared with the dimensions of the charge; the results from a series of bare charges of 70/30 T.N.T./C.E. are quoted below as an illustration of the agreement of the practical results with theory.

Each charge was in the form of a cylinder whose length was equal to its diameter, and each was exploded on the ground with its axis vertical. A list of the sizes and weights of the charges is as follows:

diameter of charge in.	weight of charge lb.
4	2¾
7	15
9	32
10	44
12	76

Blast pressure records were taken at six distances from each charge, the gauges being hung about 4 ft. above the ground.

Reproductions of the blast pressure curves from some of the charges are given in figure 11.

Figure 12 shows the variation of the maximum excess pressure with distance (in charge diameters) from the various charges. Owing to the way in which excess pressure varies with distance it is convenient to plot on logarithmic scales, and this has accordingly been done. Figure 13 is a similar plot but with linear scales, of the maximum values of the negative pressure or suction. In figure 14 the positive durations, divided by the appropriate charge diameters, have been plotted against distance in charge diameters, and in figure 15 the positive hydrostatic impulses have been graphed in the same manner.

The maximum positive pressures and the impulses are seen to lie fairly well on single curves. The maximum negative pressures and the positive durations show greater variability but can also be represented by single curves. The variability of the experimental points in figure 13 is accounted for by the difficulty of measuring accurately the small deflexions representing the suctions, and, in figure 14, is largely accounted for by irregularities in the pressure-time curves, due probably to reflected or diffracted waves from slight elevations or depressions in the ground.

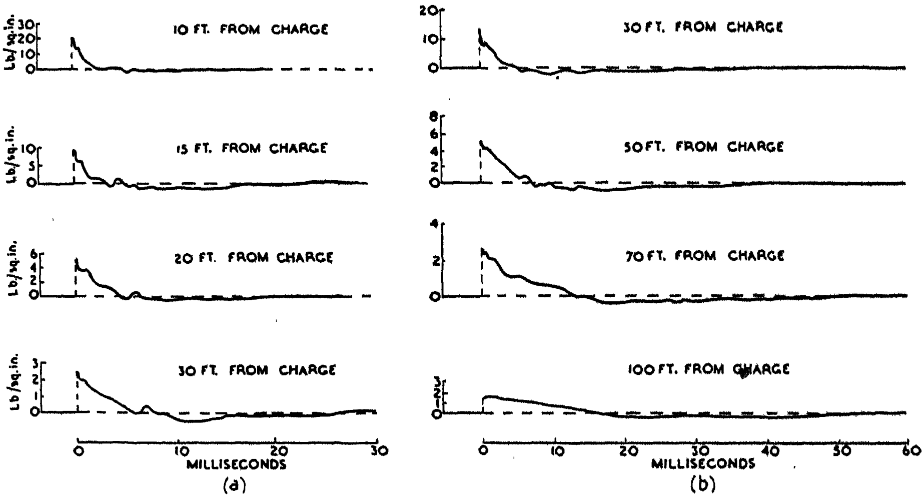


FIGURE 11. Pressure-time curves from (*a*) $2\frac{3}{4}$ lb. T.N.T./C.E. charge, and (*b*) 44 lb. T.N.T.-C.E. charge.

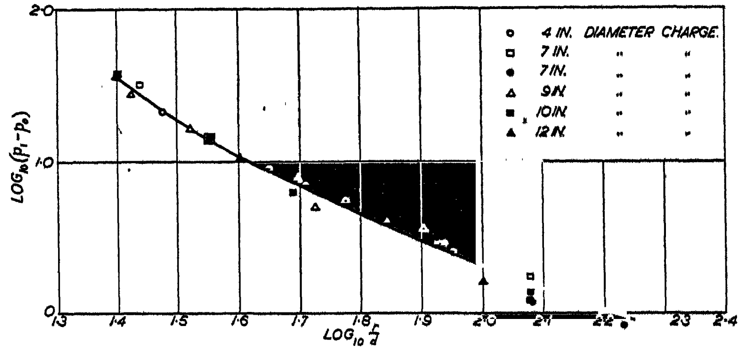


FIGURE 12. Maximum excess pressures at various distances from cylindrical T.N.T./C.E. charges.

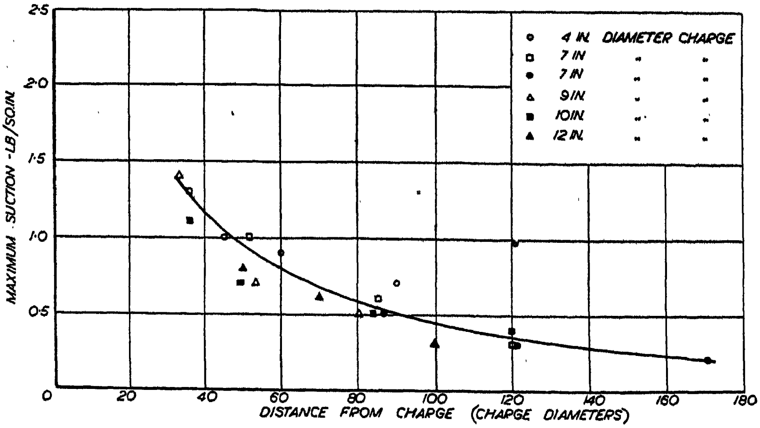


FIGURE 13. Maximum suction from T.N.T./C.E. charges.

The scale relationship in bomb testing. The scale relationship is of great importance in bomb testing, since it enables predictions of the theoretically attainable performance of new bombs to be made from the results of tests with geometrically similar bombs already in service, thus providing a standard by which the performance of a new bomb may be judged. In practice it is found that, owing to irregularities in the pressure-time curves from bombs, the maximum pressures do not scale as accurately as the impulses; since, however, the impulses are, in any case, of greater importance in determining the damage produced by a bomb than are the maximum pressures, the assessment of bomb performance is based mainly on a con-

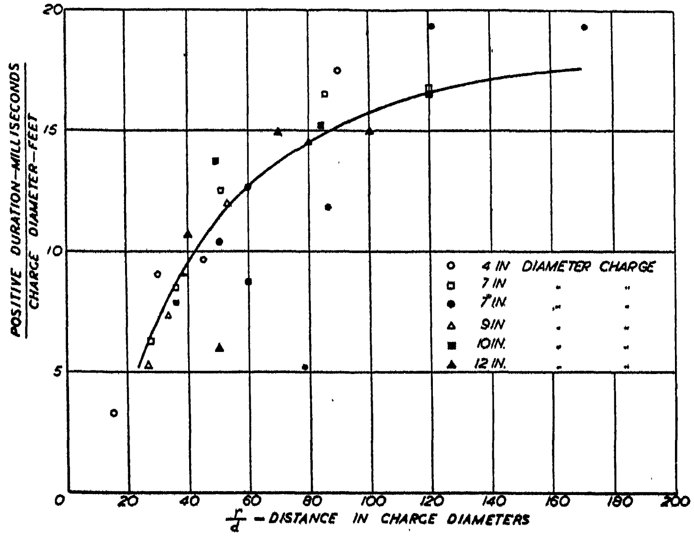


FIGURE 14. Duration of positive pressure from T.N.T./C.E. charges.

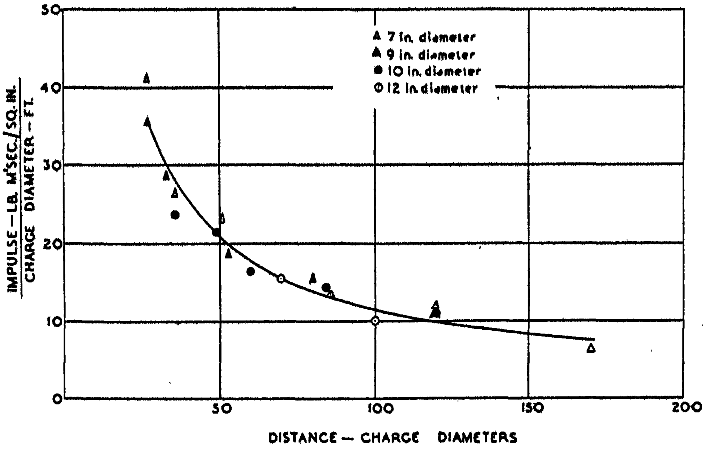


FIGURE 15. Positive impulses from cylindrical T.N.T./C.E. charges.

sideration of the hydrostatic impulses. An example of the scaling of the impulses from a series of bombs is given in figure 16, in which the results from German S.C. bombs weighing 250, 500 and 1000 kg. are plotted. In view of the fact that the bombs were not exactly similar the agreement is very good.

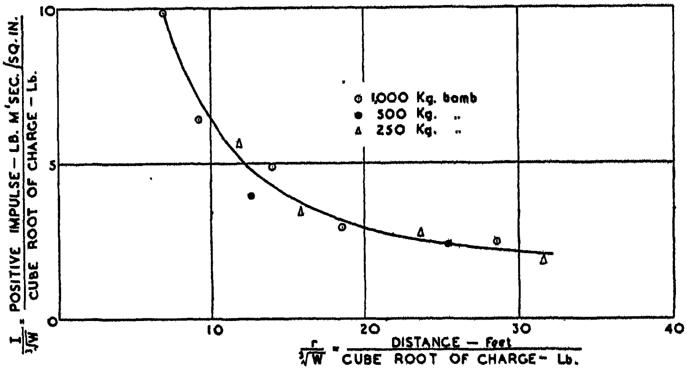


FIGURE 16. Positive impulses from German S.C. bombs.

An approach may be made to many problems of bomb design by means of tests on a model scale, with consequent saving of time and labour. A question which has recently received first investigation in this way is the influence of case thickness on the blast performance of bombs. The charges, each consisting of about 8½ lb. of 60/40 R.D.X./T.N.T., were enclosed in cylindrical steel or brass cases with flat ends, the thickness of the ends being equal to that of the case. The length of the cavity containing the charge was in all cases twice its diameter. One end of each case was screwed, to allow the charge to be inserted, and had a small hole drilled through it to allow an electric detonator to be inserted. The charges were initiated by exploders consisting of about 2½ oz. of tetryl. For comparison with the cased charges, the series also included two bare charges of the explosive. Particulars of the weights of the charges and cases and of the number of each kind fired are given in table 1, including the charge/weight ratio, that is, the ratio of the weight of charge to the total weight of case plus charge, expressed as a percentage.

Each charge was fired with its axis vertical, and its base resting on a steel plate 3 ft. square by ½ in. thick, placed on the ground, the object of the plate being to eliminate the energy loss due to cratering of the ground. Nine piezo-electric gauges were placed at distances of 10 to 50 ft., six along one line radiating from the bomb, and three along a second line at right angles to the first.

Curves showing the variation of positive hydrostatic impulse with distance from the charge for each charge/weight ratio are reproduced in

TABLE 1. PARTICULARS OF CONTAINERS

charge		empty weight		charge weight		charge/weight ratio
		lb.	oz.	lb.	oz.	%
bare charge	1	—	—	8	6½	100
	2	—	—	8	6½	100
cased charge	1	1	9¼	8	8½	84·4
	2	5	2½	8	5¾	61·9
	3	25	8¾	8	7½	24·9
	4	80	11¼	8	7	9·5
	5	148	15	8	7¼	5·4
	6	1	9¼	8	8½	84·4
	7	5	3	8	5¾	61·8
	8	25	5	8	7½	25·0
	9	80	1¾	8	7½	9·5
	10	148	15	8	8¾	5·4

figure 17. From these curves, the ratios of the positive impulses from the cased charges to the positive impulses from the bare charges at various distances have been determined, and are tabulated in table 2. It will be observed that for any particular charge/weight ratio, the positive impulse is substantially the same fraction of the bare charge impulse at all distances from the cased charge. A single curve can therefore be drawn, for the particular explosive used, giving the fractional reduction of impulse, for constant charge weight, as a function of charge/weight ratio. This has been done in figure 18.

The application of these results to actual bombs, which differ in shape from the simple cylindrical form employed in these tests, will not be discussed here, but it may be stated that, when allowance has been made for the difference in shape, the results have been found to agree well with those recorded in bomb tests.

TABLE 2. RATIOS OF POSITIVE IMPULSES FROM CASED CHARGES TO THOSE FROM BARE CHARGES

charge/ weight ratio	ratio of positive impulse from cased charge to that from bare charge					
%	10 ft.	15 ft.	20 ft.	30 ft.	50 ft.	mean value
84·4	0·870	0·863	0·865	0·848	0·831	0·855
62	0·700	0·692	0·698	0·677	0·636	0·681
25	0·465	0·479	0·478	0·476	0·424	0·465
5·4	0·380	0·399	0·400	0·395	0·351	0·385

The velocity and rate of decay of the blast wave. Comparisons have been made (1) of the observed velocity of a blast wave with that calculated from the observed values of maximum pressure, and (2) of the observed rate of

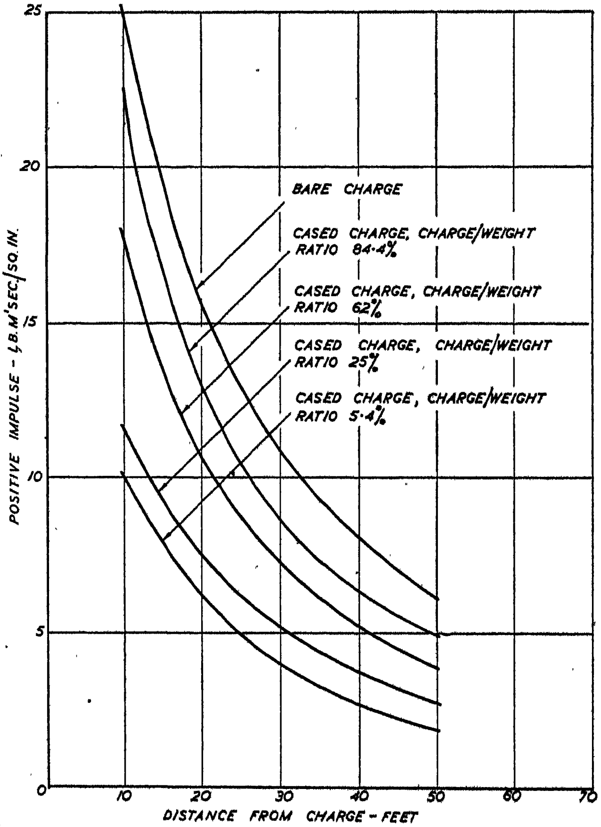


FIGURE 17. Positive impulses from cased 60/40 R.D.X./T.N.T. charges of various charge/weight ratios.

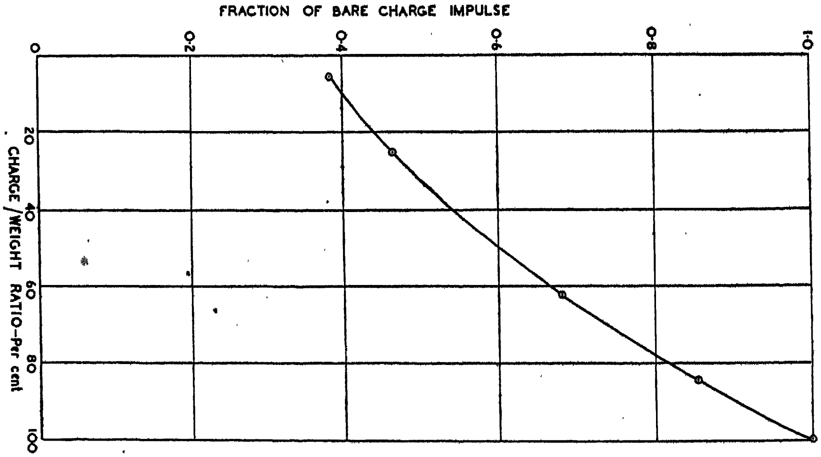


FIGURE 18. Positive impulses from cased charges of various charge/weight ratios as fractions of the bare charge impulse.

decay of a blast wave with that calculated from the observed pressure-time curves. The results used were those obtained in the tests with T.N.T./C.E. charges described in a previous section.

The mean velocities of the blast wave over certain intervals can be obtained from the records by measuring the time interval between the arrival of the shock wave at the different gauges. The mean velocities may also be calculated, since there is a definite relationship, expressed by the Rankine-Hugoniot formula, between the velocity of a shock wave advancing into still air and the pressure immediately behind the shock surface. This formula has been expressed by Professor G. I. Taylor (see ref. (3)) in the form

$$\frac{p-p_0}{p_0}=\frac{b^2-a^2}{a^2}\,\frac{2\gamma}{\gamma+1},$$

where p = maximum pressure in the blast wave,
p_0 = atmospheric pressure,
b = velocity of the shock wave,
a = velocity of sound in undisturbed air,
γ = the ratio of the specific heats.

Over the range $0 < p < 5$ atmospheres γ may be taken (see ref. (4)) as 1·405, and the expression becomes

$$\frac{p-p_0}{p_0}=\frac{1}{0{\cdot}856}\,\frac{b^2-a^2}{a^2}.$$

Thus when the maximum pressure at a point is known, the velocity of the shock wave can be calculated. In figure 19 the variation of blast velocity with distance from the charge has been calculated for the T.N.T./C.E. charges, using the mean curve of maximum excess pressure (figure 12). For comparison with the observed mean velocities the mean velocity over

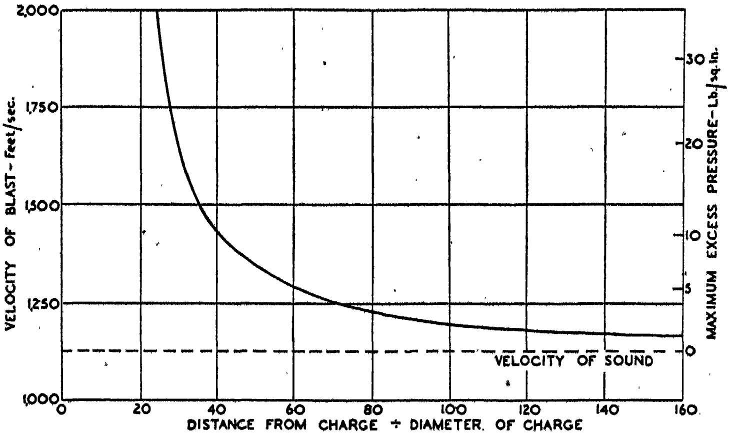


FIGURE 19. Velocity of shock wave from cylindrical T.N.T./C.E. charges calculated from pressures observed at various distances.

the same intervals must be obtained from the calculated velocities at various distances. If τ is the time taken by the shock wave to travel from r_0 to r, then $\tau = \int_{r_0}^{r} \frac{dr}{b}$, and the mean velocity $b_m = (r - r_0)/\tau$. This may be evaluated graphically.

Table 3 gives a comparison between the calculated and observed values of the mean velocity for the T.N.T./C.E. charges. The agreement between the two sets of values is good.

TABLE 3. COMPARISON OF OBSERVED AND CALCULATED VALUES FOR VELOCITY OF BLAST FROM T.N.T./C.E. CHARGES

charge	distance from charge ft.	mean velocity (ft./sec.) observed	mean velocity (ft./sec.) calculated from pressure
$2\frac{3}{4}$ lb. (4 in.)	5– 10	2140	2170
	10– 15	1930	1830
	20– 30	1200	1250
15 lb. (7 in.)	15– 20	1830	1770
	30– 50	1270	1270
	50– 70	1210	1210
32 lb. (9 in.)	25– 40	1440	1430
	40– 60	1260	1270
	60–100	1220	1200
44 lb. (10 in.)	41– 50	1290	1300
	50– 70	1220	1270
	70–100	1160	1200
	41–100	1200	1260
76 lb. (12 in.)	25– 40	1630	1610
	40– 50	1430	1390
	50– 70	1310	1310
	70–100	1150	1220
	40–100	1260	1280

When the velocity of the shock wave at any distance from the charge is known a calculation of the rate of increase of duration of the positive part of the blast wave with distance from the charge can be made, since the point behind the wave-front at which the excess pressure is zero travels with very approximately the velocity, a, of sound. Consequently, if the duration at any distance r_0 is known, the duration at any other distance r can be determined.

If t_0 is the duration at a distance r_0, the duration t_b at a distance r is given by

$$t_b = t_0 + \frac{r - r_0}{a} - \int_{r_0}^{r} \frac{1}{b}\, dr.$$

To make the calculation applicable to all sizes of charge, the duration is expressed in seconds/charge diameter, and the distances in units of charge diameter.

The curve in figure 14 which is quite a good representation of the experimental observations was calculated in this way, the value of t_b/d being taken as $7{\cdot}2 \times 10^{-3}$ at a distance of 30 charge diameters.

Calculations of the rate of decay of blast waves have been made by Taylor (see ref. (5)) and by Penney (see ref. (6)). In the Ministry of Home Security Note No. R.C. 39 on the Propagation and Decay of Blast Waves, Professor G. I. Taylor has shown that when the maximum pressure is not very high $[(p-p_0)/p_0 \not> 0{\cdot}4]$ and if the decrease in pressure behind the shock wave is approximately linear the rate of decay of maximum pressure in a spherical shock wave can be expressed approximately by the equation

$$\frac{dz}{dr}+\frac{z}{r}+\frac{\gamma+1}{4\gamma}\frac{z^2}{L}=0 \tag{1}$$

where $z=\dfrac{p-p_0}{p_0}$,

r = distance from charge,

L = length of the positive phase of the blast wave.

The formula applies to all sizes of charge if r and L are measured in units of charge diameter. If L can be regarded as constant, equation (1) can be integrated giving

$$z=\frac{z_0 r_0}{r}\,\frac{1}{1+\dfrac{\gamma+1}{4\gamma}\dfrac{z_0 r_0}{L}\log_e\dfrac{r}{r_0}},$$

where z_0 is the value of z at $r=r_0$.

Taking $\gamma = 1{\cdot}405$, this can be written

$$z=\frac{z_0 r_0}{r}\left(1+0{\cdot}984\frac{z_0 r_0}{L}\log_{10}\frac{r}{r_0}\right)^{-1}. \tag{2}$$

L is not, in fact, constant but increases with distance from the explosion. Formula (2), however, can be used over any distance which is small enough for L not to increase appreciably over that distance. Therefore, if values of Z_0 and L at a distance r_0 are assumed, the value of z at a distance r can be calculated from equation (2), if $r-r_0$ is small. The value of L at r can then be found by using the Rankine-Hugoniot formula to calculate the mean velocity b_m of the shock wave between r_0 and r, and deducing the increase in L while the shock wave is travelling from r_0 to r from the formula

$$L-L_0=(r-r_0)\left(1-\frac{a}{b_m}\right). \tag{3}$$

Using these values of r, z and L, equation (2) can then be employed again to calculate the value of z at another distance and so on.

The calculation has been carried out taking 55 charge diameters, where the maximum pressure is 5·9 lb./sq.in., as the starting point. The results are given in table 4. The calculated values are in fairly good agreement with those derived from the mean curve of figure 12, but decrease rather less rapidly with distance than do the observed pressures.

TABLE 4. COMPARISON OF CALCULATED AND OBSERVED MAXIMUM PRESSURES FROM T.N.T./C.E. CHARGES

distance from charge in charge diameters	maximum pressure (lb./sq.in.) calculated	observed
55	5·9	5·9
60	5·1	4·9
65	4·5	4·3
70	4·0	3·8
75	3·6	3·4
80	3·3	3·0
90	2·8	2·5
100	2·4	2·1
120	1·9	1·6
140	1·6	1·3
160	1·4	1·0

The paper has been prepared by permission of the Department of Scientific and Industrial Research and the other Departments interested in the work. It was written following a suggestion by the Civil Defence Research Committee of the Ministry of Home Security that a paper giving a description of the apparatus used in work for them and some results of the work should be prepared. The subject-matter includes data obtained in the course of investigations made for the Ministry of Supply and Ministry of Home Security as well as for the Department of Scientific and Industrial Research itself. The authors wish to record their indebtedness to the Assistant Director, Road Research Laboratory (Dr A. H. Davis) under whose immediate supervision the work was conducted, and to Mr W. L. Cowley, who supervised the work in its early stages.*

REFERENCES

(1) *Jber. chem. tech. Reichsanst.* 1922–7.
(2) A. R. Bryant, unpublished paper.
(3) Ministry of Home Security, Civil Defence Research Committee, Paper R.C. 39 (unpublished). G. I. Taylor.
(4) Ministry of Home Security, Civil Defence Research Committee, Paper R.C. 214, by W. G. Penney and D. R. Davies (unpublished).
(5) Loc. cit.
(6) Ministry of Home Security, Civil Defence Research Committee, Paper R.C. 260 by W. G. Penney (unpublished).

---

* The paper is submitted by permission of the Director-General of Scientific Research and Development of the Ministry of Supply and the Chief Adviser of the Ministry of Home Security, Research and Experiments Department, and with the consent of the Director of the Road Research Laboratory, Department of Scientific and Industrial Research.

Elastic constants of diamond

By S. Bhagavantam and J. Bhimasenachar

From the Department of Physics, Andhra University

(*Communicated by Sir C. V. Raman, F.R.S.—Received* 6 *October* 1945)

Employing the wedge method which has been developed by the authors, the elastic constants of diamond have been determined. The individual values are $c_{11} = 9\cdot5$, $c_{12} = 3\cdot9$, $c_{44} = 4\cdot3 \times 10^{12}$ dynes/cm.^{2}. Force constants derived from the observed spectroscopic data in respect of diamond account very satisfactorily for the results.

1. INTRODUCTION

A detailed study of the elastic behaviour of crystals is of great value for understanding the nature of the solid state. Although diamond is of special interest from this point of view, its individual elastic constants have so far not been determined presumably because of the experimental difficulties in such determinations. Adams (1921) and Williamson (1922) obtained the bulk modulus for diamond by piezometric methods. Pisharoty (1940) measured the Young's modulus for an octahedral plate of diamond by the usual bending methods.

In this paper the individual elastic constants of diamond, which are determined with the help of a new method, are given. The method has been developed by us and depends on the measurement of ultrasonic velocities in crystal plates of known orientation.*

2. THE WEDGE METHOD

Associated with a given direction of propagation in a crystal plate there are three acoustic waves, one of which may be called the quasi-longitudinal and the other two quasi-torsional. The velocities of these waves depend upon the elastic constants of the crystal. If a crystal plate of thickness d and infinite lateral dimensions is taken and a sound wave of frequency f is started at one surface, on reaching the second surface, it is partly reflected and partly transmitted. The reflected wave in conjunction with the incident wave forms a standing wave if $v = 2df$, where v is the velocity of propagation of sound. Under such circumstances the plate can be said to resonate to the frequency f. From what has been said above it is clear that a crystal plate generally resonates to three frequencies, f_1, f_2 and f_3. The plate can then be regarded as performing thickness-longitudinal or thickness-transverse vibrations according as the frequency to which it resonates is essentially a longitudinal or a transverse one. Alternatively, the plate may be regarded as a medium capable of transmitting sound waves possessing those frequencies with the highest efficiency.

---

* For details of the method and its application to other substances, reference may be made to *Proc. Indian Acad. Sci.* **20**, 298 (1944). A preliminary report of some results regarding diamond was published in *Nature*, **154**, 546 (1944).

A determination of such frequencies permits us to evaluate the characteristic values of v for the thickness direction. These values enable us to calculate the corresponding elastic constants, a number of such determinations with plates of different orientations being usually sufficient for evaluating all the independent elastic constants in any particular crystal. For plates of finite size edge effects set in, but it may be stated here that such effects have been found to be well within the experimental limits of error in our present investigations.

For determining the transmission frequencies characteristic of a given plate, a suitably cut and silvered quartz or tourmaline wedge is used. The wedge in a high-frequency circuit generates continuously varying ultrasonic frequencies, and one of these wedges is laid on a crystal plate whose orientation is known and the frequency of the electrical circuit is varied till, at a certain value, the sound wave is transmitted to a maximum extent. This particular frequency is located by allowing the sound waves transmitted through the plate to pass into a liquid and observing positions of maximum intensity in respect of the usual Debye-Sears diffraction pattern. In practice, it has been found that the setting for maximum transmission is quite sharp and capable of being reset to within 1 or 2 %. The transmission frequency is then measured with an accurate wave meter.

The transmission maxima thus recorded have to be sorted out and assigned to their appropriate modes. In an ideal case one would not expect transverse modes to be communicated to a liquid medium so as to cause Debye-Sears patterns, but work in this laboratory (Bhagavantam & Suryanarayana 1945; Bhagavantam & Bhimasenachar 1945) had shown that such modes can be excited both in piezoelectric and non-piezoelectric crystal plates and communicated to liquids in the form of consequential longitudinal strains produced by coupling due to edge effects or other inherent causes. Accordingly, thickness-longitudinal modes generally show intense Debye-Sears patterns, while the thickness-transverse modes show relatively weaker ones. Such a distinction furnishes a useful criterion by which the transmission maxima can be sorted out and associated with their corresponding modes.

3. ELASTIC CONSTANTS OF DIAMOND

In these investigations one (112) plate and three (111) plates whose orientations were checked by X-ray methods have been used. The measurements are given in table 1, c' of the last two columns stands for the effective elastic constant and is equal in each case to $4d^2f^2\rho$.

From the average values of c'

$$\tfrac{1}{2}(c_{11}+c_{12}+2c_{44}) = 11{\cdot}0, \quad \tfrac{1}{3}(c_{11}+2c_{12}+4c_{44}) = 11{\cdot}5, \quad \tfrac{1}{3}(c_{11}-c_{12}+c_{44}) = 3{\cdot}3,$$

giving $$c_{11} = 9{\cdot}5,\ c_{12} = 3{\cdot}9,\ c_{44} = 4{\cdot}3\times 10^{12}\,\text{dynes/cm.}^2.$$

TABLE 1. ELASTIC VIBRATION FREQUENCIES OF DIAMOND PLATES

plate number	orientation	thickness (mm.) d	density ρ	transmission maximum f (Mcyc.)	intensity	mode	c' value	c' expression
1	(112)	0·83	3·52	10·66	strong	longitudinal	11·03	$\frac{1}{2}(c_{11}+c_{12}+2c_{44})$
2	(111)	1·42	3·52	6·35	strong	longitudinal	11·45	$\frac{1}{3}(c_{11}+2c_{12}+4c_{44})$
3	(111)	0·79	3·50	11·48	strong	longitudinal	11·52	$\frac{1}{3}(c_{11}+2c_{12}+4c_{44})$
3	(111)	0·79	3·50	6·13	very faint	transverse	3·28	$\frac{1}{3}(c_{11}-c_{12}+c_{44})$
4	(111)	0·87	3·51	10·42	strong	longitudinal	11·53	$\frac{1}{3}(c_{11}+2c_{12}+4c_{44})$
4	(111)	0·87	3·51	5·55	faint	transverse	—	—
4	(111)	0·87	3·51	11·14	very faint	transverse overtone	3·29	$\frac{1}{3}(c_{11}-c_{12}+c_{44})$

(Units of c' are 10^{12} dynes/cm.^{2}.)

4. LATTICE THEORY OF ELASTIC CONSTANTS

The earliest attempt in the direction of giving a theory of elasticity from the atomic standpoint was that of Cauchy (1828). In recent years Born (1914) developed the theory in some detail and showed that the elastic constants of a crystal can in general be calculated if the force fields are known. Further attention was paid to the particular case of diamond in the work of Nath (1935) and of Narasimharao (1943). With a little effort, the following expressions can be obtained for the elastic constants in the case of diamond:*

$$c_{11} = \frac{1}{3d}[K_1 + 12K_\alpha + 12K_3],$$

$$c_{12} = \frac{1}{3d}[K_1 - 6K_\alpha + 6K_3],$$

$$c_{44} = \frac{1}{3d}\left[\frac{72K_1K_\alpha}{4K_1+32K_\alpha} + 6K_3\right].$$

d is the side of the unit cube. K_1 and K_3 are the force constants referring to any pair of atoms which constitute the nearest neighbours and next nearest neighbours respectively. K_α is the force constant referring to a variation in the angle between any two valence bonds which meet at an atom. A satisfactory evaluation of these force constants could be effected only recently after a thorough investigation was made of the vibration spectrum (Nayar 1942; Anna Mani 1944) and the normal oscillations (Bhagavantam 1943) of this crystal. If the diamond structure is regarded as built up of units with sixteen atoms per cell, it may be expected to exhibit eight normal modes of oscillation. Identifying (Dayal 1944) the observed values of Nayar and Anna Mani with the theoretically possible normal modes, we get the following force constants:

$$K_1 = 3{\cdot}14, \quad K_\alpha = 0{\cdot}197, \quad K_3 = 0{\cdot}39\times10^5 \text{ dynes/cm.}$$

---

* Details of evaluation are of a routine nature and are not given here, as they may be obtained from the work of Born, Nath and others quoted above.

X-ray determinations give us 3·56 A for d. Using these values in the formulae given above, calculate the elastic constants and thus

$$c_{11} = 9{\cdot}5, \quad c_{12} = 4{\cdot}0, \quad c_{44} = 4{\cdot}4 \times 10^{12} \text{ dynes/cm.}^2.$$

These compare very well with the values

$$c_{11} = 9{\cdot}5, \quad c_{12} = 3{\cdot}9, \quad c_{44} = 4{\cdot}3 \times 10^{12} \text{ dynes/cm.}^2,$$

which are the experimentally determined ones.

5. DISCUSSION OF RESULTS

The bulk modulus $K = \frac{1}{3}(c_{11} + 2c_{12})$ comes out as $5{\cdot}8 \times 10^{12}$ dynes/cm.^{2} if the values of c_{11} and c_{12} obtained in this investigation are used. This result compares favourably with Williamson's value of $5{\cdot}6 \times 10^{12}$. Adams reported $6{\cdot}3 \times 10^{12}$ for the bulk modulus, and this is rather high. Pisharoty obtained a value of $5{\cdot}5 \times 10^{12}$ for the Young's modulus with a (111) plate. This is very low compared with $9{\cdot}3 \times 10^{12}$, a value which should be expected for such a plate on the basis of our determinations.

In conclusion the authors desire to acknowledge their indebtedness to Sir C. V. Raman, Kt., F.R.S., for the readiness with which he lent the diamonds for these investigations from his collection. One of us (J. B.) would like to express his thanks to the authorities of the Andhra University for the grant of the Raman Fellowship which enabled him to work on the problem.

REFERENCES

Adams 1921 *J. Wash. Acad. Sci.* **11**, 45.
Anna Mani 1944 *Proc. Indian Acad. Sci.* **19**, 231.
Bhagavantam 1943 *Proc. Indian Acad. Sci.* **18**, 251.
Bhagavantam & Bhimasenachar 1945 *Nature*, **156**, 23.
Bhagavantam & Suryanarayana 1945 *Nature*, **155**, 171.
Born 1914 *Ann. Phys.* **44**, 605.
Cauchy 1828 *Exercise de Mathematique.*
Dayal 1944 *Proc. Indian Acad. Sci.* **19**, 224.
Narasimharao 1943 M.Sc. Thesis, Andhra University.
Nath 1935 *Proc. Indian Acad. Sci.* **1**, 333, 841.
Nayar 1942 *Proc. Indian Acad. Sci.* **15**, 293.
Pisharoty 1940 *Proc. Indian Acad. Sci.* **12**, 208.
Williamson 1922 *J. Franklin Inst.* **193**, 491.

Algebra related to elementary particles of spin $\frac{3}{2}$

BY B. S. MADHAVARAO, V. R. THIRUVENKATACHAR AND K. VENKATACHALIENGAR
Department of Mathematics, Central College, Bangalore

(*Communicated by H. J. Bhabha, F.R.S.—Received* 18 *October* 1945)

The algebra generated by the four matrices β_μ occurring in the relativistic wave equation of a particle of maximum spin n on the basis of the commutation rules for these matrices obtained previously by one of the authors has been investigated. Auxiliary quantities η_μ satisfying the equations (5) are introduced. These η_μ are given as polynomials in β_μ. With the help of these, further auxiliary quantities $\xi_\mu = \eta_\mu \beta_\mu$ are defined. It is shown that for half odd integral spin, the ξ's and η's form two mutually commuting sets of symbols of which the η's satisfy the same commutation rules as the Dirac matrices. This proves that the algebra in the case of half odd integral spin is the direct product of the Dirac algebra and an associated ξ-algebra. For the special case of maximum spin $\frac{3}{2}$ the ξ-algebra has been studied in detail, and it is shown that this algebra has just three representations of orders 1, 4, 5 such that $1^2 + 4^2 + 5^2 = 42 =$ rank of the algebra. Explicit representations are given in the non-trivial cases of orders 4 and 5. The 4-row representation of the ξ's gives a representation of the β's of order 16 which is likely to be of importance in connexion with Bhabha's new theory of the proton.

1. INTRODUCTION

One of us (B.S.M.) has considered (1942*a*, referred to here as I) the question of deriving commutation rules for the matrices β_μ appearing in the relativistic wave equation of elementary particles of arbitrary spin in the form‡

$$\partial_\mu \beta_\mu \psi + \chi \psi = 0, \tag{1}$$

and shown that this problem can be solved in the general case by postulating that the spin operator $t_{\mu\nu} = is_{\mu\nu}$ satisfies the equation (I (9*a*))

$$t_{\mu\nu} = (\beta_\mu \beta_\nu - \beta_\nu \beta_\mu) \equiv (\beta_\mu, \beta_\nu). \tag{2}$$

The general commutation rules, valid for all spins, can then be written (I (11*a*)) as

$$(\beta_\mu, t_{\nu\rho}) = (\beta_\mu, (\beta_\nu, \beta_\rho)) = \delta_{\mu\nu}\beta_\rho - \delta_{\mu\rho}\beta_\nu. \tag{3}$$

The special cases of spins $\frac{3}{2}$ and 2 were considered in I and the restricted forms of (3), on the further assumption that the eigen-values of $s_{\mu\nu}$ for a particle of spin n are $n, n-1, n-2, \ldots, -n+1, -n$, were also given there for these cases (I (26) and (34)).

The imposition of condition (2), while it solves the problem of deriving commutation rules, has also very far-reaching consequences in that it makes the wave equation (1) itself of fundamental importance in obtaining properties of elementary particles. Equations of the form (1) for which condition (2) is satisfied can be considered as natural generalizations, to general spin values, of the corresponding equations for the well-known cases of the electron and the meson. Bhabha has recently considered§ (1945 *a, b, c*) the implications of the assumption (2), and set up a general theory of relativistic wave equations of the type (1) wherein he has shown that the problem of finding all irreducible equations of the form (1) can be connected with that of finding all irreducible representations of the Lorentz group in five dimensions. He has further shown that on this theory, a particle of spin n must appear with n different values of the rest mass[‡] if n is an integer, and $n+(\frac{1}{2})$ values if n is half an odd integer, the higher values being simple rational multiples of the lowest value.

---

‡ The notation adopted in this paper is the same as that of Kemmer (1939) with χ denoting the rest mass instead of his κ.

§ We wish to thank Professor Bhabha for allowing us to see a manuscript of his paper (1945 *b*) before publication.

On the basis of his theory Bhabha has considered (1945*c*, p. 261) in particular, the two possible equations of the form (1) for a particle of maximum spin $\frac{3}{2}$, and indicated that the equation given by one of the representations of the Lorentz group in five dimensions denoted by $R_5(\frac{3}{2}, \frac{1}{2})$ may possibly describe the behaviour of the proton. The degree of this representation, viz. 16, and an explicit expression for the same have also been derived by him.

We have independently investigated the algebra generated by the symbols β_μ and 1 (unit element) governed by the commutation rules for the β_μ, and shown that, in the case of half odd integral spin, this algebra is the direct product of the Dirac algebra, and another associated algebra which we call a ξ-algebra. In particular, for the case of spin $\frac{3}{2}$ the corresponding ξ-algebra is of rank 42, and the Dirac algebra being of rank 16, the original algebra is of rank $16 \times 42 = 672$. The circumstance of this algebra being a direct product simplifies the work to a great extent, and makes it unnecessary to enumerate the 672 linearly independent elements generated by the β_μ and 1. It is shown that the ξ-algebra has three irreducible representations of orders 1, 4, 5 respectively, $(42 = 1^2+4^2+5^2)$, so that for the original algebra of spin $\frac{3}{2}$ there are also three irreducible representations of orders (4×1), (4×4), (4×5), i.e. 4, 16, 20. Of these the first refers to the well-known representation of the α-matrices of Dirac.

The commutation rules for spin $\frac{3}{2}$ are likely to prove useful in investigations relating to the above theory of Bhabha for the proton, and the associated ξ-algebra which is easier to handle than the original β-algebra might be used in calculations. We have obtained representation matrices for the ξ-algebra both of orders 4 and 5, such that the matrices are real and symmetric and thus Hermitian. A table of spurs of the elements of the basis of the ξ-algebra in both the representations is also appended.

2. THE AUXILIARY QUANTITIES η_μ, ξ_μ-ALGEBRA FOR HALF ODD INTEGRAL SPIN AS A DIRECT PRODUCT

The commutation rules satisfied by the β_μ are given by (3), and the further assumption that the eigen-values of the spin operator $s_{\mu\nu}$ for a particle of spin n are $n, n-1, \ldots, -n+1, -n$, or that $s_{\mu\nu}$ satisfies the characteristic equation

$$(s^2_{\mu\nu}-n^2)(s^2_{\mu\nu}-\overline{n-1}^2)\ldots = 0$$

[‡] See also Madhavarao (1942*b*, 1945).

leads, in virtue of (2), to the result that β_μ also satisfies the same characteristic equation

$$(\beta_\mu^2-n^2)(\beta_\mu^2-\overline{n-1}^2)\ldots=0, \tag{4}$$

the last factor being β_μ or $(\beta_\mu^2-1/4)$ according as n is an integer or half an odd integer. This is an immediate consequence of the relations

$$(\beta_\mu,\beta_\nu)=t_{\mu\nu};\quad (\beta_\mu,t_{\mu\nu})=\beta_\nu;\quad (t_{\mu\nu},\beta_\nu)=\beta_\mu$$

which follow from (2) and (3) (Bhabha 1945*a*; Madhavarao 1942*b*, 1945).

We now proceed to investigate the algebra generated by the symbols β_μ satisfying the commutation rules (3) and (4). We first of all establish the existence of a symbol η_μ satisfying the relations

$$\left.\begin{aligned}\eta_\mu^2&=1,\\ \eta_\mu\beta_\nu+\beta_\nu\eta_\mu&=0\quad(\mu\neq\nu),\\ \eta_\mu\beta_\mu&=\beta_\mu\eta_\mu\quad\text{(no summation)},\end{aligned}\right\} \tag{5}$$

Before proceeding to prove this, we might remark that the introduction of η_μ is suggested by considering the possibility of passing from the wave equation (1) to the adjoint wave equation

$$\partial_\mu\psi^\dagger\beta_\mu-\chi\psi^\dagger=0 \tag{1†}$$

with the definition $\psi^\dagger=i\psi^*\eta_4$. This is possible if we assume, as is compatible with (3) and (4), that the β_μ are Hermitian‡ (in analogy with the Dirac matrices), if $\eta_4^2=1$, $\eta_4\beta_4=\beta_4\eta_4$, and $\eta_4\beta_k+\beta_k\eta_4=0$ $(k=1,2,3)$ which suggest the relations (5).

As indicated by (4), the minimal equation for β_μ is a polynomial of degree $2n+1$, and we take $\eta_\mu=f(\beta_\mu)$ as a polynomial of degree $2n$ in β_μ, an assumption which is compatible with the relations (5). These relations further enable us to determine the coefficients of this polynomial.

Taking β_μ diagonal in an irreducible representation, the commutation rule (3) in the form

$$(\beta_\mu,(\beta_\mu,\beta_\nu))=\beta_\nu\quad(\mu\neq\nu),$$

leads to the result that the matrix for β_ν has all its elements equal to zero except those in the diagonals above and below the principal diagonal. With this form for β_ν, the relation $\eta_\mu\beta_\nu+\beta_\nu\eta_\mu=0$ shows, since the representation is irreducible, that

$$f(n)=-f(n-1)=f(n-2)=-f(n-3)=\ldots.$$

Also $\eta_\mu^2=1$ gives, since β_μ is taken diagonal and $\eta_\mu=f(\beta_\mu)$,

$$\{f(n)\}^2=\{f(n-1)\}^2=\ldots=1.$$

---

‡ That this is so has been proved by Bhabha on general considerations (1945*c*, p. 245).

Thus $f(n)$ satisfies the conditions

$$f(n) = -f(n-1) = f(n-2) = \ldots = 1 \quad (\text{or } -1),$$

i.e.
$$f(n-r) = (-1)^r \quad (r = 0, 1, 2, \ldots, 2n).$$

Since $f(x)$ is of degree $2n$ in x, it is uniquely determined by the above condition and can be written down at once, by using Lagrange's interpolation formula, in the form

$$f(x) = \frac{(x-n)(x-n+1)\ldots(x+n-1)(x+n)}{(2n)!} \sum_{r=0}^{2n} \binom{2n}{r} \frac{1}{(x-n+r)}. \tag{6}$$

For the case of half odd integral spin $n-\frac{1}{2}$, this gives

$$f(x) = \frac{2x(x^2-1/4)(x^2-9/4)\ldots(x^2-\overline{2n-1}^2/4)}{(2n-1)!} \sum_{r=1}^{n} \binom{2n-1}{n-r} \frac{1}{(x^2-\overline{2r-1}^2/4)}, \tag{6a}$$

while for integral spin n

$$f(x) = \frac{(x^2-1^2)(x^2-2^2)\ldots(x^2-n^2)}{(n!)^2} + \frac{2x^2(x^2-1^2)\ldots(x^2-n^2)}{(2n)!} \sum_{r=1}^{n} \binom{2n}{n-r} \frac{1}{(x^2-r^2)}. \tag{6b}$$

Equations (6a) and (6b) show that η_μ contains only odd powers of β_μ in the half odd integral, and only even powers in the integral case. This fact, together with $\eta_\mu\beta_\nu + \beta_\nu\eta_\mu = 0$, shows that

$$\left.\begin{aligned} \eta_\mu\eta_\nu + \eta_\nu\eta_\mu &= 0 \quad \text{for half odd integral spin,} \\ \eta_\mu\eta_\nu - \eta_\nu\eta_\mu &= 0 \quad \text{for integral spin,} \end{aligned}\right\} \tag{7}$$

i.e. η_μ and η_ν *commute or anti-commute according as the spin is integral or half odd integral.*

The η_4 introduced here is the same as the metric operator D of Bhabha (1945c (14)) given by $D = e^{i\pi\alpha^0}$ as the representative in five dimensions of the transformation of the Lorentz group reversing the axes 1, 2, 3 and interpreted as a rotation in the 04 plane through the angle π. Here similarly we can interpret η_μ as the rotation in the μ5 plane through an angle π.

We next introduce a second auxiliary symbol ξ_μ defined by

$$\xi_\mu = \eta_\mu\beta_\mu = \beta_\mu\eta_\mu \quad (\text{no summation}). \tag{8}$$

Using $\eta_\mu^2 = 1$, this leads immediately to

$$\xi_\mu\eta_\mu = \eta_\mu\xi_\mu = \beta_\mu. \tag{9}$$

Finally from $\eta_\mu\beta_\nu + \beta_\nu\eta_\mu = 0$ $(\mu \neq \nu)$ and (7), we have

$$\left.\begin{aligned} \xi_\mu\eta_\nu - \eta_\nu\xi_\mu &= 0 \quad \text{for half odd integral spin,} \\ \xi_\mu\eta_\nu + \eta_\nu\xi_\mu &= 0 \quad \text{for integral spin,} \end{aligned}\right\} \tag{10}$$

i.e. ξ_μ and η_ν *commute or anti-commute according as the spin is half odd integral or integral.*

We tabulate below the several algebraic relations between the β_μ, η_μ and ξ_μ:

$$\left.\begin{array}{c|c}
\multicolumn{2}{c}{\eta_\mu^2 = 1} \\
\multicolumn{2}{c}{\eta_\mu\beta_\nu + \beta_\nu\eta_\mu = 0} \\
\multicolumn{2}{c}{\eta_\mu\beta_\mu = \beta_\mu\eta_\mu = \xi_\mu} \\
\multicolumn{2}{c}{\eta_\mu\xi_\mu = \xi_\mu\eta_\mu = \beta_\mu} \\
\multicolumn{2}{c}{\xi_\mu^2 = \beta_\mu^2} \\
\eta_\mu\eta_\nu + \eta_\nu\eta_\mu = 0 & \eta_\mu\eta_\nu - \eta_\nu\eta_\mu = 0 \\
\eta_\mu\xi_\nu - \xi_\nu\eta_\mu = 0 & \eta_\mu\xi_\nu + \xi_\nu\eta_\mu = 0 \\
\text{for half odd integral spin} & \text{for integral spin}
\end{array}\right\} \tag{11}$$

The proof that in the case of half odd integral spin the algebra is a direct product is now immediate. Consider any expression which is a product of the powers of the several β_μ. Substituting $\beta_\mu = \eta_\mu\xi_\mu = \xi_\mu\eta_\mu$, this expression separates out, in virtue of

$$\xi_\mu\eta_\mu = \eta_\mu\xi_\mu, \quad \xi_\mu\eta_\nu - \eta_\nu\xi_\mu = 0,$$

into the product of two parts, one expressed in terms of the ξ's alone, and the other in terms of the η's alone. Further, the same substitution reduces the commutation rules (3) and (4) into rules involving the ξ's only, the η's falling out by using the relations $\eta_\mu^2 = 1$ and $\eta_\mu\eta_\nu + \eta_\nu\eta_\mu = 0$. These last relations however show that the η_μ's with their powers and products and the unit element 1 form a sub-algebra identical with the Dirac algebra, say D_η and another generated by the ξ's which might be called a ξ-algebra and denoted by A_ξ. Thus

$$A = D_\eta \times A_\xi. \tag{12}$$

3. THE ξ-ALGEBRA FOR THE CASE OF SPIN $\frac{3}{2}$

For the particular case of spin $\frac{3}{2}$, the characteristic equation (4) reduces to

$$\beta_\mu^4 - \tfrac{5}{2}\beta_\mu^2 + \tfrac{9}{16} = 0, \tag{13}$$

and this and (3) constitute the set of commutation rules for the β_μ. These can be written in such a way that they contain terms consisting of products of four β's, on the one hand, and connect these with terms containing products of two and zero β's on the other. Such types of rules can be derived from (13) by taking Poisson brackets with the $t_{\mu\nu}$ successively, and using (2) and (3). There are five such types of rules depending on the number of equal and unequal indices appearing in them. All the types including (13) can be derived from a general commutation rule: (I (26))

$$\begin{aligned}
2(\beta_\mu\beta_\nu\beta_\rho\beta_\epsilon + \beta_\mu\beta_\epsilon\beta_\rho\beta_\nu + \beta_\nu\beta_\rho\beta_\epsilon\beta_\mu + \beta_\epsilon\beta_\rho\beta_\nu\beta_\mu) \\
= 3(\beta_\mu\beta_\nu + \beta_\nu\beta_\mu)\,\delta_{\rho\epsilon} + 3(\beta_\mu\beta_\epsilon + \beta_\epsilon\beta_\mu)\,\delta_{\nu\rho} \\
+ (\beta_\rho\beta_\epsilon + \beta_\epsilon\beta_\rho)\,\delta_{\mu\nu} + (\beta_\nu\beta_\rho + \beta_\rho\beta_\nu)\,\delta_{\mu\epsilon} \\
+ (\beta_\nu\beta_\epsilon + \beta_\epsilon\beta_\nu)\,\delta_{\mu\rho} + (\beta_\mu\beta_\rho + \beta_\rho\beta_\mu)\,\delta_{\nu\epsilon} \\
- \tfrac{3}{2}(\delta_{\mu\nu}\delta_{\rho\epsilon} + \delta_{\mu\rho}\delta_{\nu\epsilon} + \delta_{\nu\rho}\delta_{\mu\epsilon}).
\end{aligned} \tag{14}$$

The individual types of rules are given in I, but for our present purposes, we write only two of them, corresponding to $\mu \neq \nu = \rho = \epsilon$, and μ, ν, ρ, ϵ unequal, i.e.

$$4(\beta_\mu^3\beta_\nu+\beta_\nu\beta_\mu^3) = 7(\beta_\mu\beta_\nu+\beta_\nu\beta_\mu) \tag{15}$$

and

$$\beta_\mu\beta_\nu\beta_\rho\beta_\epsilon+\beta_\mu\beta_\epsilon\beta_\rho\beta_\nu+\beta_\nu\beta_\rho\beta_\epsilon\beta_\mu+\beta_\epsilon\beta_\rho\beta_\nu\beta_\mu = 0. \tag{16}$$

The commutation rule (15) suggests immediately that the auxiliary symbol η_μ can be taken, in virtue of $\eta_\mu\beta_\nu+\beta_\nu\eta_\mu = 0$, as

$$\eta_\mu = \text{const.}\,(4\beta_\mu^3-7\beta_\mu)$$

and $\eta_\mu^2 = 1$ gives, using (13), this constant equal to $\pm\frac{1}{3}$. We therefore take

$$\eta_\mu = \tfrac{1}{3}(4\beta_\mu^3-7\beta_\mu) \tag{17}$$

which is also obtained from (6a) by putting $n = 2$ in that equation.

The auxiliary ξ_μ is given by $\xi_\mu = \eta_\mu\beta_\mu$ and, using (17) and (13), this gives $\xi_\mu = \beta_\mu^2-\frac{3}{4}$ and, since $\xi_\mu^2 = \beta_\mu^2$, we have

$$\xi_\mu^2 = \xi_\mu+\tfrac{3}{4}, \tag{18}$$

which is a very important equation since it materially simplifies the ξ-algebra.

We can now derive the commutation rules for the ξ_μ by substituting $\beta_\mu = \xi_\mu\eta_\mu = \eta_\mu\xi_\mu$ in the commutation rules for the β_μ and removing the η_μ by using the relations $\eta_\mu^2 = 1$, $\eta_\mu\eta_\nu+\eta_\nu\eta_\mu = 0$ and $\eta_\nu\xi_\mu-\xi_\nu\eta_\mu = 0$. This can be done for all the types of rules derivable from (14), but it is however remarkable that by using (18) and the general rule (3) expressed in terms of the ξ's, all these types of commutation rules are identically satisfied with the single exception of (16). The commutation rules for the ξ_μ therefore consist of (18) and (3) and (16) expressed in terms of the ξ's. Using the notation $(a, b) \equiv ab-ba$ and $[a, b] \equiv ab+ba$, (3) expressed in the ξ's gives the two rules

$$[\xi_\mu, [\xi_\mu, \xi_\nu]] = \xi_\nu \;\; (\mu \neq \nu) \quad \text{and} \quad (\xi_\mu, [\xi_\nu, \xi_\rho]) = 0 \;\; (\mu, \nu, \rho \neq), \tag{19a, b}$$

and (16) in terms of the ξ's becomes

$$\xi_\mu(\xi_\nu\xi_\rho\xi_\epsilon-\xi_\epsilon\xi_\rho\xi_\nu) = (\xi_\nu\xi_\rho\xi_\epsilon-\xi_\epsilon\xi_\rho\xi_\nu)\,\xi_\mu \tag{20}$$

with μ, ν, ρ, ϵ unequal.

(18), (19) and (20) constitute the commutation rules for the ξ's. We can write them in further abbreviated form by writing

$$\xi_{\mu\nu\rho} \equiv \xi_\mu\xi_\nu\xi_\rho \quad \text{and} \quad (\xi_{\mu\nu\rho}) \equiv (\xi_{\mu\nu\rho}-\xi_{\rho\nu\mu}),$$

and using a similar notation for products containing two or four terms. The rules then become

$$\left.\begin{aligned} \xi_\mu^2 &= \xi_\mu+\tfrac{3}{4}, \\ \xi_{\mu\nu}+\xi_{\nu\mu}+2\xi_{\mu\nu\mu} &= -\tfrac{1}{2}\xi_\mu \quad (\mu \neq \nu), \\ (\xi_{\mu\nu\rho}) = (\xi_{\nu\rho\mu}) &= (\xi_{\rho\mu\nu}) \quad (\mu \neq \nu \neq \rho), \\ \xi_\mu(\xi_{\nu\rho\epsilon}) &= (\xi_{\nu\rho\epsilon})\,\xi_\mu \quad (\mu \neq \nu \neq \rho \neq \epsilon). \end{aligned}\right\} \tag{21a–d}$$

It is now easy to enumerate the number of independent elements of the algebra A_ξ. In virtue of (21*a*) there are only twelve elements of the type $\xi_{\mu\nu}$. From (21*c*) we see that there are eight relations among the twenty-four products $\xi_{\mu\nu\rho}$ for which $\mu \neq \nu \neq \rho$, and hence there are only sixteen elements of this type. Also the elements $\xi_{\mu\nu\rho}$ for which two or more indices are equal would be enumerated under ξ_μ or $\xi_{\mu\nu}$ according to (21*a*, *b*). Similar remarks apply to $\xi_{\mu\nu\rho\epsilon}$ when two or more indices are equal. We have therefore to enumerate such terms when all the indices are different. The number of relations among the twenty-four products $\xi_{\mu\nu\rho\epsilon}$ can be enumerated by writing (21*d*) using (21*c*) in the form

$$\left.\begin{aligned} \xi_\mu(\xi_{\nu\rho\epsilon}) &= (\xi_{\nu\rho\epsilon})\,\xi_\mu \\ = \xi_\mu(\xi_{\rho\epsilon\nu}) &= (\xi_{\rho\epsilon\nu})\,\xi_\mu \\ = \xi_\mu(\xi_{\epsilon\nu\rho}) &= (\xi_{\epsilon\nu\rho})\,\xi_\mu \end{aligned}\right\} \tag{22}$$

with similar relations having ξ_ν, ξ_ρ, ξ_ϵ in place of the ξ_μ, and giving in all four sets of five relations of the type (22). It is, however, easily verified that of these four sets, any one can be obtained from the remaining three by mere addition, so that there are only fifteen relations among the $\xi_{\mu\nu\rho\epsilon}$ and hence only nine independent elements of that type.

The complete list of independent elements of A_ξ in consonance with (21) is thus given by the table:

element	number of type
I	1
ξ_μ	4
$\xi_{\mu\nu}$	12
$\xi_{\mu\nu\rho}$	16
$\xi_{\mu\nu\rho\epsilon}$	9
total number	42

(23)

Hence it follows from (12) that the number of independent elements of the original algebra A is $16 \times 42 = 672$, the Dirac algebra being of rank 16.

Let us find the elements of A_ξ which commute with all the others. We write

$$\left.\begin{aligned} \alpha &= \Sigma\xi_\mu, \\ \beta &= \Sigma\xi_{\mu\nu}, \\ \gamma &= \Sigma\xi_{\mu\nu\rho}, \\ \delta &= \Sigma\xi_{\mu\nu\rho\epsilon}, \end{aligned}\right\} \tag{24}$$

the indices being unequal in each of the summations. By successively postulating commutability with each of the ξ_μ, we find that the following expressions have the desired property:

$$I \text{ (the unit element)}, \quad M = \alpha - \beta, \quad N = 2\gamma - \delta. \tag{25}$$

Any further expressions which might be shown to commute with all the forty-two elements can be proved to be linear combinations of 1, M, N. We also note that the elements α, β, γ, δ form a commutative sub-algebra with the following relations:

$$\left.\begin{array}{rl|rl} & & 2\alpha\beta = 2\beta\alpha &= 3\alpha+2\gamma \\ & & 4\beta\gamma = 4\gamma\beta &= 15\alpha+2\gamma \\ \alpha^2 &= \alpha+\beta+3 & 2\gamma\alpha = 2\alpha\gamma &= 5\beta+2\gamma+2\delta \\ 2\beta^2 &= 2\beta+2\delta+9 & \alpha\delta = \delta\alpha &= \gamma \\ 4\gamma^2 &= 15\alpha+5\beta-4\gamma-4\delta+45 & 2\beta\delta = 2\delta\beta &= 5\beta-4\delta \\ 4\delta^2 &= -20\beta+16\delta+45 & 4\gamma\delta = 4\delta\gamma &= 15\alpha-4\gamma \end{array}\right\} \tag{26}$$

from which follow the relations

$$\left.\begin{aligned} M^2 &= \tfrac{9}{2}-M-N, \\ MN &= -\tfrac{15}{2}M, \\ N^2 &= \tfrac{225}{4}. \end{aligned}\right\} \tag{27}$$

The primitive idempotent elements of the centre of the ξ-algebra e_1, e_2, e_3 satisfying

$$e_1+e_2+e_3 = 1, \quad e_1e_2 = e_2e_3 = e_1e_3 = 0,$$

are given by

$$\left.\begin{aligned} 30e_1 &= 2N+15, \\ 48e_2 &= 6M-2N+15, \\ -80e_3 &= 10M+2N-15. \end{aligned}\right\} \tag{28}$$

As the algebra A_ξ has the unit element

$$1 = e_1+e_2+e_3,$$

where

$$e_ie_k = \delta_{ik}e_k \quad (i, k = 1, 2, 3),$$

it follows (see van der Waarden 1931, II, p. 161, ex. 4) that A_ξ is the direct sum of the left ideals $A_\xi e_1$, $A_\xi e_2$ and $A_\xi e_3$, and hence from the principal theorem relating to semi-simple rings (ibid. p. 156) we conclude that A_ξ is semi-simple.

Therefore the number of irreducible representations is equal to the rank of the centrum which is *three* in our case. If n_1, n_2, n_3 be the degrees of these representations we have

$$n_1^2+n_2^2+n_3^2 = 42, \tag{29}$$

the algebra A_ξ being of rank 42. The equation (29) has the unique solution in integers given by $n_1 = 4$, $n_2 = 5$, $n_3 = 1$, and thus the algebra has three and only three inequivalent irreducible representations D_1, D_4, D_5 of degrees 1, 4 and 5 respectively.

The left ideals $A_\xi e_1$, $A_\xi e_2$ and $A_\xi e_3$ are of ranks 16, 25 and 1 respectively as could be easily verified. Thus the representations of order 1, 4, 5 are generated respectively by the idempotent elements e_3, e_1, e_2.

4. REPRESENTATIONS, AND CORRESPONDING SPURS

We now proceed to give the explicit representations of degrees four and five of A_ξ. The one-dimensional representation in which every ξ_μ is represented by $-\frac{1}{2}$ leads just to the Dirac matrices for the representation of A, and need not be specially considered here.

Fourth order. The eigen-values of ξ_μ are given by (21*a*) as $-\frac{1}{2}$ and $\frac{3}{2}$, and taking ξ_4 diagonal, we will show that consistent with the commutation rules (21) the form of ξ_4 is uniquely determined by taking the eigen-value $-\frac{1}{2}$ repeated thrice and $\frac{3}{2}$ appearing only once.

Taking ξ_4 diagonal, we can write it in the form

$$\xi_4 = \left(\begin{array}{c|c} -\frac{1}{2}E_{mm} & 0 \\ \hline 0 & \frac{3}{2}E_{nn} \end{array}\right),$$

where E_{mm} denotes the $m \times m$ unit matrix, and $m+n = 4$. Let

$$\xi_1 = \left(\begin{array}{c|c} A_{mm} & B_{mn} \\ \hline C_{nm} & D_{nn} \end{array}\right),$$

where A_{ij} denotes a matrix of i rows and j columns. From the commutation rule (21*b*) or (19*a*)

$$\xi_1 = [\xi_4, [\xi_4, \xi_1]] \quad \text{and} \quad \xi_4 = [\xi_1, [\xi_4, \xi_1]]$$

and, using the above matrix forms for ξ_4 and ξ_1, we get the conditions

$$D_{nn} = 0, \tag{29a}$$

$$C_{nm}B_{mn} = \tfrac{3}{4}E_{nn}, \tag{29b}$$

$$A_{mm} = 2B_{mn}C_{nm} - \tfrac{1}{2}E_{mm}, \tag{29c}$$

in deriving which use has also been made of $\xi_1^2 = \xi_1 + \frac{3}{4}$.

From (29*b*) we obtain that $C_{nm}B_{mn}$ must be of rank n; hence $m \geqslant n$, i.e. in a D_4, $-\frac{1}{2}$ must appear at least twice as an eigen-value of ξ_4. We prove that it must occur thrice, for if it occurs twice, $m = n = 2$, and we have from (29*b*) and (29*c*)

$$C_{nn} = \tfrac{3}{4}B_{nn}^{-1} \quad \text{and} \quad A_{nn} = \tfrac{3}{2}B_{nn}B_{nn}^{-1} - \tfrac{1}{2}E_{nn} = E_{nn},$$

i.e. ξ_1 is of the form

$$\xi_1 = \left(\begin{array}{c|c} E_{22} & X \\ \hline \frac{3}{4}X^{-1} & 0 \end{array}\right).$$

Arguing similarly with ξ_2, ξ_3 in place of ξ_1 these would be of the form

$$\xi_2 = \left(\begin{array}{c|c} E_{22} & Y \\ \hline \frac{3}{4}Y^{-1} & 0 \end{array}\right), \quad \xi_3 = \left(\begin{array}{c|c} E_{22} & Z \\ \hline \frac{3}{4}Z^{-1} & 0 \end{array}\right).$$

Using the commutation rule (21*c*) or (19*b*), we have $(\xi_4, [\xi_1, \xi_2]) = 0$ which gives $X + Y = 0$. Similarly $(\xi_4, [\xi_2, \xi_3]) = 0$ and $(\xi_4, [\xi_3, \xi_1]) = 0$ give $Y + Z = 0$ and

$Z+X=0$. Thus $X=Y=Z=0$, but this contradicts the fact that the minimal equation of ξ_μ is $\xi_\mu^2-\xi_\mu-\frac{3}{4}=0$. Hence $m=n=2$ is not possible and the only possible case is $m=3$, $n=1$ so that $-\frac{1}{2}$ appears thrice and $\frac{3}{2}$ only once in ξ_4.

We notice that this choice is in consonance with the representation of order 16 given by Bhabha (1945 *c*, p. 261) and denoted in the five-dimensional notation as $R_5(\frac{3}{2},\frac{1}{2})$. As shown there, the eigen-value of α^0 (corresponding to our β_4) taken in the diagonal form as $\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}$ which occur 2, 6, 6, 2 times respectively.

With
$$\xi_4=\left(\begin{array}{c|c} -\frac{1}{2}E_{33} & 0 \\ \hline 0 & \frac{3}{2} \end{array}\right)$$

we can now write ξ_k $(k=1,2,3)$, using (29), in the form

$$\xi_k=\left(\begin{array}{c|c} 2B_{31}^{(k)}C_{13}^{(k)}-\frac{1}{2}E_{33} & B_{31}^{(k)} \\ \hline C_{31}^{(k)} & 0 \end{array}\right),$$

where $C_{13}^{(k)}B_{31}^{(k)}=\frac{3}{4}$. As we wish to take a real, symmetric (Hermitian) representation we take $C_{13}^{(k)}=(B_{31}^{(k)})'$ (the dash denoting the transpose) and obtain

$$\xi_k=\left(\begin{array}{c|c} 2P_kP'_k-\frac{1}{2} & P_k \\ \hline P'_k & 0 \end{array}\right), \tag{30}$$

where P_k is a one-column matrix
$$P_k=\begin{pmatrix} p_k \\ q_k \\ r_k \end{pmatrix}, \tag{31}$$

and P'_k its transpose, satisfying the relations

$$P'_kP_k=\tfrac{3}{4}, \quad P'_iP_k=\tfrac{1}{4} \quad (i\neq k), \ (i,k=1,2,3), \tag{32}$$

the latter being a consequence of $(\xi_4,[\xi_i,\xi_k])=0$.

We make a particular choice of the p, q, r, and derive the representation:

$$\left.\begin{aligned}
&\xi_1=\left(\begin{array}{c|c|c|c} c & \frac{1}{2} & . & c \\ \hline \frac{1}{2} & -s & . & s \\ \hline . & . & -\frac{1}{2} & . \\ \hline c & s & . & . \end{array}\right) \qquad
\xi_2=\left(\begin{array}{c|c|c|c} -s & . & \frac{1}{2} & s \\ \hline . & -\frac{1}{2} & . & . \\ \hline \frac{1}{2} & . & c & c \\ \hline s & . & c & . \end{array}\right) \\
&\xi_3=\left(\begin{array}{c|c|c|c} -\frac{1}{2} & . & . & . \\ \hline . & c & \frac{1}{2} & c \\ \hline . & \frac{1}{2} & -s & s \\ \hline . & c & s & . \end{array}\right) \qquad
\xi_4=\left(\begin{array}{c|c|c|c} -\frac{1}{2} & . & . & . \\ \hline . & -\frac{1}{2} & . & . \\ \hline . & . & -\frac{1}{2} & . \\ \hline . & . & . & \frac{3}{2} \end{array}\right)
\end{aligned}\right\} \tag{33}$$

where
$$s=\sin\frac{\pi}{10}=\frac{\sqrt{5}-1}{4}; \quad c=\cos\frac{2\pi}{10}=\frac{\sqrt{5}+1}{4}.$$

In the above representation ξ_4 is taken as diagonal. We can also derive a non-diagonal representation of our ξ's by making use of the following considerations. In virtue of (2) the β_μ along with the $t_{\mu\nu}$ form ten infinitesimal transformations I^{KL} (see Bhabha 1945*b*, §2, and 1945*c*, p. 244) of the nucleus of an irreducible representation of the Lorentz group in five dimensions with K and L taking values from 1 to 5 (or in Bhabha's notation, from 0 to 4). The division of these ten transformations into two sets of four and six is achieved by merely writing $I^{\mu 5} = \beta_\mu$, i.e. denoting the set I^{15}, I^{25}, I^{35}, I^{45} by β_1, β_2, β_3, β_4. A completely equivalent way of the subdivision of the I^{KL} would be to take I^{12}, I^{13}, I^{14}, I^{15} in place of the β_μ's. In the notation of our ξ-algebra, this means that in place of ξ_1, ξ_2, ξ_3, ξ_4 as fundamental symbols generating the algebra, we might take ξ_1, $-[\xi_1, \xi_2]$, $-[\xi_1, \xi_3]$, $-[\xi_1, \xi_4]$. We now write ξ_1 in the form (30) and make a suitable choice of the p, q, r appearing therein, and thus obtain the matrix for ξ_1. For example a particular choice yields

$$\xi_1 = \left(\begin{array}{c|c|c|c} 2 & \frac{i\sqrt{5}}{2} & -\frac{i\sqrt{5}}{2} & \frac{\sqrt{5}}{2} \\ \hline \frac{i\sqrt{5}}{2} & -1 & \frac{1}{4} & \frac{i}{2} \\ \hline -\frac{i\sqrt{5}}{2} & \frac{1}{4} & -1 & -\frac{i}{2} \\ \hline \frac{\sqrt{5}}{2} & \frac{i}{2} & -\frac{i}{2} & \end{array}\right). \tag{33'}$$

The matrices for ξ_2, ξ_3, ξ_4 are then obtained so as to be in consonance with the fact that the ξ_1, $-[\xi_1, \xi_2]$, $-[\xi_1, \xi_3]$ and $-[\xi_1, \xi_4]$ satisfy the same commutation rules as ξ_1, ξ_2, ξ_3, ξ_4. The representation thus obtained is non-diagonal.

By taking the direct product with the Dirac matrices η_μ, we can obtain from (33) and (33′) two equivalent types of representation matrices of order 16 for the β_μ in one of which β_4 is diagonal and not so in the other. By a permutation of rows and columns, we have been able to show that the matrices of the former type for β_k can be brought to the form given by Bhabha (1945*c*, p. 251, (35)) in the case of general spin (with α^0 diagonal). For the case of spin $\frac{3}{2}$ in particular, Bhabha has further given (1945*c*, p. 261, (75), (76)) representation matrices for the β_k expressed in terms of the u and v matrices introduced by Dirac (1936) and Fierz (1939). Writing these matrices in numerical form using the expressions given by Fierz (ibid. Appendix) we can derive the corresponding 16×16 matrices of the representation in which β_4 is not diagonal. It is also possible to show that the second type of matrices which we obtain from (33′) would be equivalent, but for permutation of rows and columns, to the above matrices derived from those given by Bhabha.

Fifth order. Proceeding as in the case of the fourth order representation, and taking ξ_4 diagonal, we deduce that, consistent with the commutation rules, the form of ξ_4 is uniquely determined by taking the eigen-value $-\frac{1}{2}$ as appearing thrice and $\frac{3}{2}$

twice. This is also in consonance with Bhabha's result (1945 *c*, p. 259) that the eigenvalues of β_4 in the diagonal form are $\frac{3}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$, $-\frac{3}{2}$ appearing 4, 6, 6, 4 times respectively. Thus we write

$$2\xi_4 = \operatorname{diag}(-1^3, 3^2),$$

and find that the ξ_k $(k = 1, 2, 3)$ are given by the same expressions as (30) where

$$P_1 = \frac{\sqrt{3}}{2}\begin{pmatrix} p & a \\ q & b \\ p-q & a-b \end{pmatrix}, \quad P_2 = \frac{\sqrt{3}}{2}\begin{pmatrix} -p & -a \\ -q & -b \\ p-q & a-b \end{pmatrix}, \quad P_3 = \frac{\sqrt{3}}{2}\begin{pmatrix} -p & -a \\ q & b \\ q-p & b-a \end{pmatrix}, \tag{34}$$

with a, b, p, q, subject only to the conditions

$$\left.\begin{aligned} a^2+b^2-ab &= p^2+q^2-pq = \tfrac{1}{2}, \\ 2(ap+bq) &= aq+bp. \end{aligned}\right\} \tag{35}$$

We make the particular choice

$$p = q = \frac{1}{\sqrt{2}}, \quad a = -b = \frac{1}{\sqrt{6}},$$

so as to obtain the matrices Hermitian (in fact real and symmetric). This gives

$$\left.\begin{aligned}
2\xi_1 &= \left(\begin{array}{c|c|c|c|c} 1 & 1 & 1 & \sqrt{(\frac{3}{2})} & (1/\sqrt{2}) \\ \hline 1 & 1 & -1 & \sqrt{(\frac{3}{2})} & -(1/\sqrt{2}) \\ \hline 1 & -1 & 1 & \cdot & (\sqrt{2}) \\ \hline \sqrt{(\frac{3}{2})} & \sqrt{(\frac{3}{2})} & \cdot & \cdot & \cdot \\ \hline (1/\sqrt{2}) & -(1/\sqrt{2}) & (\sqrt{2}) & \cdot & \cdot \end{array}\right), \quad
2\xi_2 = \left(\begin{array}{c|c|c|c|c} 1 & 1 & -1 & -\sqrt{(\frac{3}{2})} & -(1/\sqrt{2}) \\ \hline 1 & 1 & 1 & -\sqrt{(\frac{3}{2})} & (1/\sqrt{2}) \\ \hline -1 & 1 & 1 & \cdot & (\sqrt{2}) \\ \hline -\sqrt{(\frac{3}{2})} & -\sqrt{(\frac{3}{2})} & \cdot & \cdot & \cdot \\ \hline -(1/\sqrt{2}) & (1/\sqrt{2}) & (\sqrt{2}) & \cdot & \cdot \end{array}\right) \\
2\xi_3 &= \left(\begin{array}{c|c|c|c|c} 1 & -1 & 1 & -\sqrt{(\frac{3}{2})} & -(1/\sqrt{2}) \\ \hline -1 & 1 & 1 & \sqrt{(\frac{3}{2})} & -(1/\sqrt{2}) \\ \hline 1 & 1 & 1 & \cdot & -(\sqrt{2}) \\ \hline -\sqrt{(\frac{3}{2})} & \sqrt{(\frac{3}{2})} & \cdot & \cdot & \cdot \\ \hline -(1/\sqrt{2}) & -(1/\sqrt{2}) & -(\sqrt{2}) & \cdot & \cdot \end{array}\right), \quad
2\xi_4 = \left(\begin{array}{c|c|c|c|c} -1 & & & & \\ \hline & -1 & & & \\ \hline & & -1 & & \\ \hline & & & 3 & \\ \hline & & & & 3 \end{array}\right)
\end{aligned}\right\} \tag{36}$$

Similar remarks apply here, as regards non-diagonal representations, as for the fourth order.

We next consider the spurs of the independent elements of the basis of the algebra A. Since the spurs of all the basis elements of the Dirac algebra D_η except I are zero, it follows from (12) that the only elements of the basis of A which have non-vanishing spurs are those which have non-vanishing spurs among the ξ's. We therefore list below the spurs of the elements of the basis of the ξ-algebra in the three irreducible representations:

$$\begin{array}{lccc} \text{representation} & \text{1-row} & \text{4-row} & \text{5-row} \\ \text{spur } I & 1 & 4 & 5 \\ \text{spur } \xi_{\mu} & -\frac{1}{2} & 0 & \frac{3}{2} \\ \text{spur } \xi_{\mu\nu} & \frac{1}{4} & 0 & -\frac{3}{4} \\ \text{spur } \xi_{\mu\nu\rho} & -\frac{1}{8} & \frac{1}{2} & -\frac{5}{8} \\ \text{spur } \xi_{\mu\nu\rho\epsilon} & \frac{1}{16} & -\frac{1}{4} & \frac{5}{16} \end{array} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\end{array}}\right\}. \qquad (37)$$

By means of these, we can write down immediately the spurs of all combinations of the β's by using the algebraic relations (11) and the fact that all elements of D_{η}, except I have zero spur. Thus for example, the spur calculations relating to products of powers of the several β_{μ} given by Booth & Wilson (1940, pp. 498–9) can be carried out with less labour in our case with the aid of the ξ's than in the case of the 10-row representation of the meson algebra. Similarly for spur calculations relating to Bhabha's new theory of the proton, it would not be necessary to use the 16×16 β-matrices, but only the 4×4 ξ-matrices.

We wish to thank Professor Bhabha for helpful discussions.

References

Bhabha 1945*a* *Curr. Sci.* **14**, 89–90.
Bhabha 1945*b* In course of publication in *Rev. Mod. Phys.*
Bhabha 1945*c* *Proc. Indian Acad. Sci.* **21**, 241–264.
Booth & Wilson 1940 *Proc. Roy. Soc.* A, **175**, 483–518.
Dirac 1936 *Proc. Roy. Soc.* A, **155**, 447–459.
Fierz 1939 *Helv. Phys. Acta*, **12**, 3–37.
Kemmer 1939 *Proc. Roy. Soc.* A, **173**, 91–116.
Madhavarao 1942*a* *Proc. Indian Acad. Sci.* **15**, 139–147.
Madhavarao 1942*b* *J. Mys. Univ.* **3**, (10), 59–63.
Madhavarao 1945 *J. Mys. Univ.* **6**, (5), 57–62.
van der Waarden 1931 *Moderne Algebra*, **2**, 156, 161.

Some thermodynamical properties of gaseous sulphur dichloride

By C. A. McDowell and E. A. Moelwyn-Hughes

(*Communicated by Professor A. R. Todd, F.R.S.—Received* 29 *October* 1945)

The fundamental vibration frequencies of sulphur dichloride have been estimated from the geometry and dimensions of the molecule. These vibration frequencies and the known molecular data have been used to calculate statistically the various thermodynamic properties of gaseous sulphur dichloride.

On account of its inherent instability, sulphur dichloride, SCl_2, has proved to be a difficult compound to investigate. Scepticism has, in fact, prevailed concerning its existence, which has not long been firmly established (Lowry, McHatton & Jones 1927; Lowry & Jessop 1929; Naik & Jadhov 1926). Calorimetric data are still largely lacking, but recent electron-diffraction investigations (Palmer 1938) make it possible to estimate theoretically the magnitude of some of its thermodynamic properties. The results of such estimates are given in this paper. They are intended as provisional values, which serve to fill a gap in the literature until they can be superseded by direct calorimetric measurements, or by calculations based on more extensive spectroscopic data.

We are primarily interested in the Helmholtz free energy, the Gibbs function and the entropy, which are derivable for an ideal gaseous system of N molecules in terms of the effective partition function, f:

Helmholtz free energy: $$(A^0 - E_0^0) = -NkT \ln f; \tag{1}$$

Gibbs function: $$(G^0 - E_0^0) = -NkT\left\{\ln f - \left(\frac{d\ln f}{d\ln V}\right)_T\right\}; \tag{2}$$

Entropy: $$S^0 = Nk\left\{\ln f + T\left(\frac{d\ln f}{dT}\right)_V\right\}. \tag{3}$$

The effective partition function, f, is related to the absolute partition function by means of the degeneracy, g_0, and the residual energy, ϵ_0, which characterize the molecule in its ground state:

$$f_{\text{absolute}} = g_0 e^{-\epsilon_0/kT} f. \tag{4}$$

Translational motion. The molecular partition function for the translational motion of a molecule in the standard ideal gas state is given by the equation

$$f^0_{\text{trans.}} = \frac{g(2\pi mkT)^{\frac{3}{2}}\, kTe}{h^3 P^0}, \tag{5}$$

where Boltzmann's constant, $k = 1{\cdot}372 \times 10^{-16}$ erg/degree; Planck's constant, $h = 6{\cdot}552 \times 10^{-27}$ erg-sec; $P^0 = 1{\cdot}014 \times 10^6$ dynes/cm.^{2}; and m, the mass of SCl_2

molecule, is $16{\cdot}982 \times 10^{-23}$ g. By analogy with OCl_2 (Mulliken 1935) the quantum degeneracy, g, is assumed to be unity for SCl_2.

Rotational motion. The partition function for the rotational motion of a triatomic angular molecule such as SCl_2 is given by the equation

$$f_{\text{rot.}} = \frac{\pi^{\frac{1}{2}}(8\pi^2kT)^{\frac{3}{2}}\,(I_A I_B I_C)^{\frac{1}{2}}}{\sigma h^3}, \tag{6}$$

where I_A, I_B and I_C are the moments of inertia of the molecule, and σ, the symmetry factor, is 2 for SCl_2. Electron diffraction data give the S-Cl distance in both SCl_2 and S_2Cl_4 as $1{\cdot}99 \pm 0{\cdot}03$ A (Palmer 1938). The covalent radius of the sulphur atom is 1·04 A (Moelwyn-Hughes 1940), and for the chlorine atom the covalent radius is 0·98 A (Pauling 1942); hence the S-Cl distance is $2{\cdot}02 \pm 0{\cdot}01$ A. Taking the S-Cl distance as 2·02 A and the valency angle as 101° (Palmer 1938), then the three moments of inertia, I_A, I_B, I_C, will be $(60{\cdot}11, 284{\cdot}2 \text{ and } 344{\cdot}3) \times 10^{-40}$ g.-cm.^{2} respectively.

Vibrational motion. The vibrational partition function for a non-linear triatomic molecule is

$$f_{\text{vib.}} = \prod_{i=1}^{i=3} (1 - e^{-h\nu_i/kT})^{-1}. \tag{7}$$

There are, unfortunately, neither infra-red nor Raman data known for SCl_2, but as this is an angular symmetrical triatomic molecule its fundamental vibrational frequencies could be calculated provided one had some knowledge of the force constants for the three fundamental types of internal motion. The force constant for the S-Cl link may be calculated from the empirical relation (Badger 1934)

$$b = \frac{C_{ij}}{(r_e - d_{ij})^3}, \tag{8}$$

b = force constant, r_e = interatomic distance, and C_{ij} and d_{ij} are constants which depend on the position of the constituent atoms in the periodic table.

For sulphur dichloride then

$$b = \frac{1{\cdot}86 \times 10^5}{(2{\cdot}02 - 1{\cdot}18)^3} = 3{\cdot}14 \times 10^5 \text{ dynes/cm.}$$

As a check on this value the following independent method of calculating the force constant can be used. Assume the potential energy ϕ of the two atoms to be represented by the equation

$$\phi = Ar^{-n} - Br^{-m}, \tag{9}$$

then it can be shown (Sutherland 1938) that the force constant, b, is given by the relation

$$b = \frac{mnD_e}{r_e^2}, \tag{10}$$

where $D_e = -\phi_e$, the minimum value of ϕ, and r_e is the corresponding value of r, the interatomic distance. From certain unpublished experiments it has been estimated that the energy of fission of the S-Cl link is 71 kcal./g.mol.; hence taking r_e as $2{\cdot}02 \times 10^{-8}$ cm. then $b = mn \times 1{\cdot}20 \times 10^4$ dynes/cm. Now it can be demonstrated from experimental information on diatomic molecules that mn for S-S is 23·7, and for Cl-Cl is 29·3. Taking mn for S-Cl as 26 then $b = 3{\cdot}12 \times 10^5$ dynes/cm., which is in good agreement with the value obtained from Badger's relation. Therefore $b = 3{\cdot}13 \times 10^5$ dynes/cm. may be accepted as a probable value for the force constant of the S-Cl link.

To determine the fundamental vibrational frequencies it is necessary to know the potential energy of the molecule as a function of the interatomic distances of the constituent atoms. In general, this function is not known, but by analogy with diatomic molecules, it may be assumed that the increase, ΔV, in the potential energy, relative to a zero value in the equilibrium state, can be written as the sum of three quadratic terms:

$$\Delta V = \tfrac{1}{2}\{b_1(\delta r_{ab})^2 + b_2(\delta r_{bc})^2 + b_\alpha(\delta\alpha)^2\}. \tag{11}$$

δr and $\delta\alpha$ denote the small differences produced in the interatomic separations (r_{ab} and r_{bc}) and in the valency angle (α), for the system displaced very slightly from the equilibrium state; and b_1, b_2 and b_α are the corresponding force constants (Cross & Van Vleck 1933; Penney & Sutherland 1936). For a symmetrical triatomic molecule like SCl_2, $r_{ab} = r_{bc} = r_e$ and $b_1 = b_2 = b$; hence equation (11) becomes

$$\Delta V = \tfrac{1}{2}\{2b(\delta r_e)^2 + b_\alpha(\delta\alpha)^2\}. \tag{12}$$

With a potential function of this form the fundamental vibrational frequencies can be shown to be given by the following equations:

$$(2\pi\nu_3)^2 = \frac{b}{m_a}\left\{1 + \frac{2m_a}{2m_a+m_b}\sin^2\left(\frac{\alpha}{2}\right)\right\}, \tag{13a}$$

$$(2\pi\nu_1)^2 + (2\pi\nu_2)^2 = \frac{b}{m_a}\left\{1 + \frac{2m_a}{2m_a+m_b}\cos^2\left(\frac{\alpha}{2}\right)\right\} + \frac{b_\alpha}{m_a r_e^2}\left\{1 + \frac{2m_a}{2m_a+mb}\sin^2\left(\frac{\alpha}{2}\right)\right\}, \tag{13b}$$

$$(2\pi\nu_1)^2\,(2\pi\nu_2)^2 \doteq \frac{bb_\alpha}{m_a^2 r_e^2}\left\{1 + \frac{2m_a}{2m_a+m_b}\right\}. \tag{13c}$$

ν_1, ν_2 and ν_3 are the fundamental vibrational frequencies; m_a is the mass of the chlorine atom; m_b is the mass of the sulphur atom; α the valency angle of SCl_2; b the force constant for the S-Cl link $= 3{\cdot}13 \times 10^5$ dynes/cm.; b_α is the force constant for the valency vibration. The value of b_α for CS_2 is $2{\cdot}2 \times 10^{-11}$ dyne-cm./radian, and for Cl_2O it is $2{\cdot}4 \times 10^{-11}$ dyne-cm./radian (Cross & Van Vleck 1933); hence b_α for SCl_2 may be taken to be $2{\cdot}3 \times 10^{-11}$ dyne-cm./radian. Substituting the respective values for the parameters in the above equations it is estimated that the wave numbers corresponding to the fundamental vibrational frequencies of SCl_2 are $\omega_1 = 404$ cm.$^{-1}$, $\omega_2 = 206$ cm.$^{-1}$ and $\omega_3 = 461$ cm.$^{-1}$. As stated above, experimental confirmation of these frequencies is lacking, but they may be compared with those observed for sulphur monochloride, S_2Cl_2, which are: 110 cm.$^{-1}$ (strong); 200 cm.$^{-1}$ (medium); 247 cm.$^{-1}$ (weak); 451 cm.$^{-1}$ (very strong) (Kohlrausch 1937).

Employing the forms of the partition functions and the data given above we have calculated from equations (1), (2), (3) and (4) the effective Helmholtz free energy ($A^0 - E_0^0$), the Gibbs function ($G^0 - E_0^0$) and entropy S^0 for 1 g.mol. of SCl_2 in the standard gas state at 1 atm. pressure; these data are recorded in the most generally useful form in table 1. In calculating the vibrational contributions we made use of the tables published by Wilson (1940).

TABLE 1. THE HELMHOLTZ FREE ENERGY, GIBBS'S FUNCTION AND ENTROPY OF SULPHUR DICHLORIDE IN THE STANDARD GAS STATE AT 1 ATM. PRESSURE

T° K	$\left(\frac{A^0-E_0^0}{T}\right)$ cal.g.mol.$^{-1}$deg.$^{-1}$	$\left(\frac{G^0-E_0^0}{T}\right)$ cal.g.mol.$^{-1}$deg.$^{-1}$	S_0 cal.g.mol.$^{-1}$deg.$^{-1}$
291·1	59·32	57·33	67·57
298·1	59·55	57·57	67·87
400	62·65	60·67	71·62
500	65·19	63·21	74·67
600	67·32	65·34	77·11
700	69·14	67·16	79·18
800	70·74	68·76	80·99
900	72·19	70·21	82·59
1000	73·53	71·55	84·07

Other thermodynamic properties of sulphur dichloride. The internal energy E, the heat content H, and the molar heat capacity C_p^0 of SCl_2 in the standard gas state have been calculated from the equations given below. The contributions of the vibrational motion to these quantities have been evaluated from Wilson's tables. The internal energy is given by the expression

$$(E^0-E_0^0) = NkT\left\{3+\sum_{i=1}^{3}\left(\frac{h\nu_i}{kT}\right)\frac{e^{-h\nu_i/kT}}{1-e^{-h\nu_i/kT}}\right\}. \qquad (14)$$

The heat content $(H^0-E_0^0) = (E^0-E_0^0)+RT$; hence the molar heat capacity $C_p^0 = \left(\frac{dH}{dT}\right)_p$ is given by the expression

$$C_p^0 = Nk\left\{4+\sum_{i=1}^{3}\left(\frac{h\nu_i}{2kT}\operatorname{cosech}\frac{h\nu_i}{2kT}\right)^2\right\}. \qquad (15)$$

TABLE 2. THE INTERNAL ENERGY, HEAT CONTENT, AND MOLAR HEAT CAPACITY OF SULPHUR DICHLORIDE IN THE STANDARD GAS STATE AT 1 ATM. PRESSURE

	cal.g.mol.$^{-1}$ deg.$^{-1}$				
T° K	$\left(\frac{E-E_0}{T}\right)_{\text{vib.}}$	$\left(\frac{E_0-E_0^0}{T}\right)_{\text{total}}$	$\left(\frac{H_0-E_0^0}{T}\right)_{\text{total}}$	$(C_p)_{\text{vib.}}$	C_p^0
291·1	2·293	8·254	10·241	4·58	12·53
298·1	2·350	8·311	10·298	4·64	12·58
400	3·008	8·969	10·956	5·16	13·10
500	3·509	9·470	11·457	5·43	13·38
600	3·820	9·781	11·768	5·59	13·53
700	4·073	10·034	12·021	5·68	13·63
800	4·282	10·234	12·230	5·74	13·69
900	4·439	10·400	12·387	5·79	13·74
1000	4·577	10·538	12·525	5·82	13·77

We wish to thank Dr W. C. Price and Dr G. B. B. M. Sutherland for supplying certain spectroscopic information. We are also indebted to the late Professor Sir Ralph Fowler, F.R.S., for his helpful criticism of this paper, and to the Chief Scientific Officer of the Ministry of Supply for permission to publish it.

REFERENCES

Badger 1934 *J. Chem. Phys.* **2**, 128.
Cross & Van Vleck 1933 *J. Chem. Phys.* **1**, 350.
Kohlrausch 1937 *Der Smekal-Raman Effekt*, p. 342. Springer.
Lowry & Jessop 1929 *J. Chem. Soc.* p. 1421.
Lowry, McHatton & Jones 1927 *J. Chem. Soc.* p. 746.
Moelwyn-Hughes 1940 *Physical chemistry*, p. 49. Cambridge.
Mulliken 1935 *J. Chem. Phys.* **3**, 512.
Naik & Jadhov 1926 *J. Indian Chem. Soc.* **3**, 360.
Palmer 1938 *J. Amer. Chem. Soc.* **60**, 2360.
Pauling 1942 *The nature of the chemical bond*, 2nd ed. p. 165.
Penney & Sutherland 1936 *Proc. Roy. Soc.* A, **156**, 654.
Sutherland 1938 *Proc. Ind. Acad. Sci.* **8**, 341.
Wilson 1940 *Chem. Rev.* **27**, 27.

The mechanism of ionospheric ionization. II

By R. v. d. R. Woolley

(*Communicated by S. Chapman, F.R.S.—Received* 18 *December* 1945)

The available mechanisms for the production of electrons in the three regions of the ionosphere are discussed with special reference to the question whether it is possible to account for the observed electron densities without supposing that the sun emits far more energy in the remote ultra-violet spectrum than would be emitted by a black body at 6000°. The contributions to electron densities made by metastable states of atoms and molecules are examined. It is concluded that the observed electron densities may be accounted for without requiring high solar energy in the ultra-violet if the effective recombination coefficient in the F_2 region is 10^{-11}. The F_2 region is supposed formed by the ionization of atomic oxygen, and the E region by the ionization of molecular oxygen. The electrons forming the F_1 region are supposed to be provided by metastable N_2 or by NO.

1. The electron density in the earth's atmosphere considered as a function of height exhibits several maxima which are observed by the reflexion of radio waves sent out from ground-level. Broadly speaking three such maxima are observed; the neighbourhood of each maximum is called a region, and the regions are named E, F_1 and F_2 in order of ascending height. The E region reflects all waves having frequencies less than the critical frequency of the E region. Waves above this frequency and below the critical frequency of the F_1 region are reflected by the F_1 region, and waves of still greater frequency are reflected by the F_2 region. Waves of frequency greater than the critical frequency of the F_2 region are not reflected at all. Since the critical frequency is proportional to the square root of the maximum electron density in the layer, the maximum electron densities ascend from E to F_2. It follows that there may well be unobservable maxima in the electron density.

The mathematical theory of the formation of a single layer is due to Chapman (1931). Suppose that for a particular absorption spectrum in the atmosphere capable of ionizing an atom or molecule the absorption coefficient $k(\nu)$ is independent of ν for all frequencies greater than ν_0 and is zero for $\nu < \nu_0$. The number of electrons produced per c.c. from n atoms (or molecules) per c.c. is nq, where

$$q = \int_{\nu_0}^{\infty} kS_\nu (h\nu)^{-1} e^{-\tau \sec z} d\nu,$$

where S_ν is the intensity of sunlight entering the atmosphere, τ is the optical depth in the atmosphere or $\tau = \int_x^\infty kn\,dx$ and z is the sun's zenith distance: x is the height measured upwards.

Now suppose that the atoms or molecules are distributed exponentially with height according to the law $n = n_0 e^{-x/H}$, where n_0 and H are constant, then $\tau = Hkn_0 e^{-x/H}$. Write $\int_{\nu_0}^{\infty} S_\nu (h\nu)^{-1} d\nu = S$. Then

$$nq = Skn_0 e^{-x/H} \exp\{-Hkn_0 \sec z\, e^{-x/H}\}.$$

Let $nq = Q$. For a given zenith distance the maximum value of Q is $Q_m(z)$ occurring at $x = x_m(z)$, where

$$x_m(z) = H \ln (Hkn_0 \sec z), \tag{1·11}$$

and if $N(z)$ is the value of n where the maximum occurs then

$$HkN(z) = \cos z. \tag{1·12}$$

Again $$Q_m(z) = Skn_0 \exp\{-x_m(z)/H - 1\} = S \cos z/(eH). \tag{1·13}$$

Writing $\xi = (x_m(z) - x)/H$, then

$$Q/Q_m(z) = \exp\{1 + \xi - e^{\xi}\}. \tag{1·14}$$

The quantities concerned may also be referred to $Q_m(0)$, the maximum value of Q when the sun is in the zenith, or $z = 0$. Then

$$x_m(0) = H \ln Hkn_0, \tag{1·21}$$

$$HkN_0 = 1, \tag{1·22}$$

$$Q_m(0) = S/(eH), \tag{1·23}$$

and writing $\zeta = (x_m(0) - x)/H$,

$$Q/Q_m(0) = \exp\{1 + \zeta - \sec z \, e\zeta\}. \tag{1·24}$$

These expressions are due to Chapman. It may be shown that they are still approximately true if $k(\nu)$ varies appreciably with ν. For example, if $k(\nu) = 0$ for $\nu < \nu_0$ and $k(\nu) = \alpha\nu^{-3}$ for $\nu > \nu_0$, these expressions are still very nearly correct in the practical case where S_ν is a black-body curve for a temperature T' and $h\nu_0/(RT') > 20$.

2. The electron density n_e can be found from Q if the absorption spectrum is the *only* method whereby electrons are produced and if the recombination coefficient α is known; for then there are αn_e^2 recombinations per c.c. per second and

$$\frac{dn_e}{dt} = Q - \alpha n_e^2.$$

Taking an average over 24 hr. it may be supposed that dn_e/dt is zero, and then $Q = \alpha n_e^2$. This expression can be reduced to a special form if the recombination is radiative, that is to say, if it is the reverse of the ionization process, for there is then a relation between the absorption coefficient k and the recombination coefficient α as was shown by Milne.

In full equilibrium at a temperature T between n_e electrons, n atoms and n^+ ions per c.c., then, according to Saha,

$$\left(\frac{n_e n^+}{n}\right)_T = K(T) = 2(2\pi mRT)^{\frac{3}{2}} h^{-3} \varpi \exp(-I/RT),$$

where I is the ionization potential and ϖ is the ratio of the statistical weights of the ion and atom. m is the mass of the electron and R, h Boltzmann's and Planck's constants respectively.

Now in full equilibrium $nq_T = \alpha' n_e n^+$ $(= \alpha n_e^2)$, and in secular equilibrium $(dn_e/dt = 0)$, $nq = \alpha' n_e n^+$. Hence in secular equilibrium

$$\frac{n_e n^+}{n} = \frac{q}{q_T} K(T).$$

The value of $qK(T)/q_T$ has been investigated by Woltjer (1925) in a special case, namely, where the radiation intensity is a fraction (less than unity) of full radiation at some temperature T' while the temperature of the atmosphere (judged by the Maxwellian velocity distribution of the molecules) is T, less than T'. Woltjer's equations have been applied to stellar atmospheres by Pannekoek (1926) and to the earth's atmosphere by Bhar (1938). The subject is discussed fully by Pannekoek in the *Handbuch der Astrophysik*, **3/1**, 288–291.

Let the intensity of the radiation be $\frac{\omega}{8\pi} B(\nu, T')$, where $B(\nu, T') = \frac{8\pi h\nu^3}{c^2} \frac{1}{e^{h\nu/RT'} - 1}$ is the intensity of full radiation at the temperature T'. Then

$$\frac{n_e n^+}{n} = \omega K(T, T') = \frac{\omega}{8\pi} K(T) \frac{\int k(\nu)\, B(\nu, T')\, (h\nu)^{-1}\, d\nu}{\int k(\nu)\, B(\nu, T)\, (h\nu)^{-1}\, d\nu}. \qquad (2{\cdot}1)$$

In particular, if $k(\nu) = 0$ for $h\nu < I$ and $k(\nu) :: \nu^{-2}$ for $h\nu > I$,

$$\omega K(T, T') = 2(2\pi m R)^{\frac{3}{2}} T^{\frac{1}{2}} T' \frac{\omega}{8\pi} \frac{\varpi}{h^3} \exp(-I/RT'). \qquad (2{\cdot}2)$$

If $k(\nu) = 0$ for $h\nu < I$ and is constant for $h\nu > I$, the right-hand side of (2·2) must be multiplied by

$$\left\{1 + 2\frac{RT'}{I} + 2\left(\frac{RT'}{I}\right)^2\right\} \Big/ \left\{1 + 2\frac{RT}{I} + 2\left(\frac{RT}{I}\right)^2\right\}. \qquad (2{\cdot}21)$$

If there are n^- negative ions per c.c. then, in place of (2·1),

$$\frac{n_e n}{n^-} = \omega K(T, T'), \qquad (2{\cdot}3)$$

with, of course, values of I and ϖ which differ from those in the positive-ion reaction. In the case of O^-, ϖ is $\frac{2}{3}$ and according to Vier & Mayer (1944) $I = 3{\cdot}07$ V. Then with $T' = 6825°$, $T = 1000°$ and $\omega = 5{\cdot}41 \times 10^{-6}$

$$\frac{n_e n}{n^-} = 3 \times 10^{12}.$$

Thus with $n = 3 \times 10^{10}$ and $n_e = 10^6$, representing conditions in the F_2 region, $n^- = 10^4$, giving one negative ion to about one hundred electrons.

3. Before attempting to identify the various layers of the ionosphere with particular ionization processes, note must be taken of a difficulty which is illustrated by the following theorem. Let there be a number of ionization processes and let

ν_n be the lowest frequency capable of ionizing the atom or molecule in the nth process, numbered downwards in the order in which the maxima occurs in the atmosphere (e.g. $F_2 = 1$, $F_1 = 2$ and $E = 3$, if these regions arise from the only processes). Let S_n be the amount of light available for the nth process. Let α_n be the effective recombination coefficient in the region where the ionization due to the nth process produces a maximum electron density. *Then in an isothermal atmosphere, if the quantities S_n/α_n form an increasing series, only one layer, the lowest, is observable.*

This is since $\alpha_n(N_e^2)_n = (Q_m(z))_n = S_n \cos z/(eH)$,

so that if S_n/α_n increases with decreasing height, $(N_e^2)_n$ increases with decreasing height and the lowest layer has the largest electron density and reflects all the radio waves capable of reflexion by any higher layer. Hence the higher layers cannot be observed.

In the actual atmosphere, the temperature probably increases with height so that H is greater for F_2 than for E. On the other hand, α decreases from about 10^{-8} at E to about 10^{-10} or 10^{-11} at F_2. With the former value (10^{-10}), αN_e^2 is roughly constant from layer to layer, and the quantities S_n must increase with increasing height in the series of observed layers.

If the solar-energy curve is at all like a black-body curve at 6000°, S_ν decreases very rapidly with increasing ν in the far ultra-violet, and a series of absorption head frequencies will give a series of values of S_n which increase rapidly with decreasing ν_n unless the absorption heads are very close together. If each layer completely cuts off the radiation capable of ionizing the next

$$S_n = \int_{\nu_n}^{\nu_{n+1}} S(\nu)\,(h\nu)^{-1}\,d\nu.$$

Values of $S'_n = 10^{-8}\beta^{-3}\int_{\nu_n}^{\nu_{n+1}} \nu^2 e^{-\beta\nu}\,d\nu$ are shown for relevant values of $\beta\nu_n$ in table 1.

TABLE 1. S'_n

$\beta\nu_n$	24	25	26	27	28
S'_n	142·3	56·7	22·5	8·93	3·53

Since for $T' = 6000°$ and $h\nu = 13{\cdot}54\,\mathrm{V}$ (the ionization potential of O), $\beta\nu = 26{\cdot}2$, successive entries in table 1 correspond to potential differences of about half a volt. A series of values of S'_n increasing with ν_n could only be obtained if successive ionization potentials closer than a quarter of a volt apart were available. In particular, the identification of two of the three layers with the reactions

$$\mathrm{O}+h\nu\ (>13{\cdot}5\,\mathrm{V})\rightarrow\mathrm{O}^{+}+e,\quad \mathrm{O}_2+h\nu\ (>12{\cdot}5\,\mathrm{V})\rightarrow\mathrm{O}_2^{+}+e,$$

is not possible if αN_e^2 is constant—unless, of course, one abandons the hypothesis that $S(\nu)$ does not decrease with increasing ν so rapidly as a black body at 6000° or so. If, however, $\alpha = 10^{-11}$ for the F_2 region, in which case αN_e^2, and, therefore, Q, is about ten times larger in both of the lower regions, S'_n for the F_2 region has only one-tenth of the value for a lower region, and from table 1, $\beta\nu = 25$ and $\beta\nu = 27$ give approximately correct values of S'_n.

According to Appleton (1937) the value of α for the F_2 region is 10^{-10}. This value depends upon afternoon minus morning differences in the F_2 critical frequencies which are attributed to recombination. According to Martyn (1945) tidal motion of charged particles is a main cause of the asymmetries in the diurnal ionization variations of the F_2 region: admitting Martyn's hypothesis, the value of α cannot be determined with any satisfaction from diurnal variations in critical frequency.

Now if $\frac{dn_e}{dt} = -\alpha n_e^2$, then $\frac{d}{dt}(\ln n_e) = -\alpha n_e$. With $n_e = 10^6$, then $-\frac{d}{dt}(\ln n_e) = 10^{-4}$ sec.$^{-1}$, or (3 hr.)$^{-1}$, if $\alpha = 10^{-10}$, and (28 hr.)$^{-1}$ with $\alpha = 10^{-11}$. According to Allen's analysis (1946) of the correlations between a variety of solar and terrestrial phenomena, the F_2 region exhibits a mean period of recombination of order 1 or 2 days. This observation supports $\alpha = 10^{-11}$. The well-known lag (of the order of a day) of F_2 critical frequencies behind magnetic storms gives further support to the lower value of α. In the remainder of this paper $\alpha = 10^{-11}$ will be used for the F_2 region.

4. In this section the conditions under which two observable regions could be formed by the ionization of O and of N_2 is examined, waiving the supposition that the solar-energy curve is that of a black body if necessary. Suppose that the atmosphere is isothermal and n_1, n_2 are the numbers of oxygen atoms and nitrogen molecules per c.c. Let k_1, k_2 be the continuous absorption coefficient of atomic oxygen and molecular nitrogen. Following Massey & Bates (1943) take k_1 to be independent of wave-lengths from 912 A downwards: and suppose k_2 to be independent of wave-lengths from 661 A (the beginning of the strong absorption of N_2) to shorter wave-lengths.

Let $\qquad S_1 = \int_{\nu_1}^{\nu_2} S_\nu (h\nu)^{-1}\,d\nu \quad$ and $\quad S_2 = \int_{\nu_2}^{\infty} S_\nu (h\nu)^{-1}\,d\nu,$

where $\qquad c{\nu_1}^{-1} = 912\,\text{A} \quad$ and $\quad c{\nu_2}^{-1} = 661\,\text{A}.$

Lastly, let $\qquad n_1 = n_1^0 e^{-x/H_1} \quad$ and $\quad n_2 = n_2^0 e^{-x/H_2}.$

Then $H_2 = \frac{16}{28} H_1$. Now

$$Q = S_1 k_1 n_1^0 e^{-x/H_1} \exp\{-H_1 k_1 \sec z\, n_1^0 e^{-x/H_1}\}$$
$$+ S_2\{k_1 n_1^0 e^{-x/H_1} + k_2 n_2^0 e^{-x/H_2}\} \exp\{-H_1 k_1 \sec z\, n_1^0 e^{-x/H_1} - H_2 k_2 n_2^0 \sec z\, e^{-x/H_2}\}.$$

Write $S_2/S_1 = s$, $k_2 n_2^0/k_1 n_1^0 = \beta$ and $H_1 k_1 n_1^0 \sec z = A$. Let $x/H_1 = \xi$. Then

$$Q/(S_1 k_1 n_1^0) = \exp\{-\xi - Ae^{-\xi}\}\,[1 + s(1 + \beta e^{-\frac{3}{4}\xi}) \exp(-\tfrac{4}{7} A\beta e^{-\frac{7}{4}\xi})]. \tag{4·1}$$

According to (4·1) two maxima of Q can occur. Reckon n_1^0 and n_2^0 at the point where one of these maxima occurs: then one maximum is at $\xi = 0$, giving

$$0 = (A-1)\{1 + s(1+\beta)^2 \exp(-\tfrac{4}{7}A\beta)\} + s\beta(\beta + \tfrac{1}{4}) \exp(-\tfrac{4}{7}A\beta). \tag{4·21}$$

Let the other maximum occur at $\xi = -h$ and write $Ae^h = A'$, $\beta e^{\frac{3}{4}h} = \beta'$. Then

$$Q = (A'-1)\{1 + s(1+\beta')^2 \exp(-\tfrac{4}{7}A'\beta')\} + s\beta'(\beta' + \tfrac{1}{4}) \exp(-\tfrac{4}{7}A'\beta'). \tag{4·22}$$

The ratio of the two maxima is given by

$$Q_0/Q_{-h} = \exp\{A' - A - h\}\{1 + s(1+\beta)\, e^{-\frac{1}{2}A\beta}\}\{1 + s(1+\beta')\, e^{-\frac{1}{2}A'\beta'}\}^{-1}. \qquad (4\cdot23)$$

From (4·21), (4·22) and (4·23) the location of the two maxima in the atmosphere can be computed, and the ratio of their intensities, given the fundamental data A, s and β. Conversely, any two layers can be examined and the values of A, s and β required to account for them. Since nothing is known about s the latter course will be pursued.

Now from (4·21) and (4·22) both A and A' (which must be positive) are less than unity, so that solutions which are well separate (e^h appreciably greater than 1), A must be small. Suppose as a first approximation that $A' = 1$. Then $A = e^{-h}$ and (4·21), (4·23) give a pair of equations for s and β. When the solution has been made it can be verified from (4·22) that $1 - A'$ is small.

First suppose that $Q_0 = 10^{-1} Q_{-h}$, which may be supposed to correspond to the F_2 and F_1 regions with $\alpha = 10^{-11}$ in the upper region. No solution (with positive values) is obtained unless

$$\frac{Q_0}{Q_{-h}} \exp\{A + h - A'\} > 1,$$

so that if $Q_0/Q_{-h} = 10^{-1}$, $h > 3\cdot265$. Thus

$$\text{for} \quad h = 3\cdot401, \quad \beta = 258 \quad \text{and} \quad s = 0\cdot074,$$

$$\text{for} \quad h = 3\cdot912, \quad \beta = 142 \quad \text{and} \quad s = 0\cdot031.$$

On the other hand, suppose that the recombination coefficient is 10^{-10} in the F_2 region so that $Q_0/Q_{-h} = 1$, then

$$\text{for} \quad h = 3, \quad \beta = 36\cdot7 \quad \text{and} \quad s = 0\cdot485,$$

$$\text{for} \quad h = 2\cdot303, \quad \beta = 19\cdot0 \quad \text{and} \quad s = 0\cdot460.$$

These figures show that β is in all cases so large that unless k_2 is very much larger than k_1 (which is not likely, as the oscillator strength of the O absorption is high) the nitrogen absorption contributes mainly to the *upper* maximum. Bhar (1938) supposes that N_2 forms the *lower* of the two regions. He is in this way able to evade the difficulty described in § 3, although he supposes that the solar energy curve is that of a black body: but his result depends on the assumption that radiation at 661 A is not absorbed by O.

For a black body at 6000°, s is of order 10^{-4}, if the absorption heads are at 912 and 661 A.

Of two layer solutions of this type, even admitting the high values of s required, it must be remarked that they demand a rather higher ratio of nitrogen to oxygen population than seems at all likely. (Parallel calculations assuming the nitrogen to be in the atomic rather than the molecular state also give high values of β.) It may also be remarked that the solutions with $Q_0/Q_{-h} = 10^{-1}$ are open to the objection that they require h greater than 3, which corresponds to a separation of about 150 km. between the layers. Under many circumstances F_2 and F_1 come much closer together than this.

5. The lines of N_2^+ are actually observed in aurorae and in the light of the night sky, and to account for their excitation Saha (1935) postulated 'with a certain amount of hesitation' that the solar spectrum exhibits the Lyman lines of H and the principal lines of He and other elements as intensely bright lines on a much feebler continuous background. Later, Saha (1937) repeated this postulate much more emphatically, concluding that solar ultra-violet light of wave-length 660 A is about a million times as intense as that given out by a black body at 6500°, and stating that it was difficult to see any alternative to the result.

To obtain *three* observable layers one must have the spectrum from 661 A onwards including the He line at 584 A about one-tenth as bright as the spectrum from 912 to 661 A, and that spectrum in turn half as bright as the spectrum capable of ionizing O_2 (at 12·5 V). The Lyman α and β lines will not ionize O_2 in its ground state.

Very bright lines are impossible if the solar atmosphere is in radiative equilibrium. Without going into great detail one can say roughly that if the bright lines are formed by the cyclic process described by Rosseland (1926) no bright line can emit more energy than a black-body spectrum appropriate to the colour temperature of the star integrated from the frequency of the line to infinity. (Admitting Rosseland's principle, any small volume of the atmosphere can add no more energy to a principal line that it absorbs in higher frequencies.) It is certainly not necessary to suppose that principal lines of a spectrum are bright at all (let alone intensely bright) merely because subordinate lines (in eruptions, such as Hα) appear bright.

On the other hand, the phenomenon of very high excitation in coronal lines has not yet been justified on any theory of radiative equilibrium, and in fact Saha (1945) has given a theory of the excitation of these lines by nuclear processes. It may therefore be felt that any objection to the postulate of bright lines based on the supposition that the solar envelope is in radiative equilibrium is not a fatal objection. However, in succeeding sections various mechanisms will be examined which may account for the observed ionosphere without invoking this postulate.

6. Suppose that the ionization of atomic oxygen in its ground state

$$O(^3P) + h\nu \rightarrow O^+ + e \tag{6·1}$$

forms a simple Chapman layer. Then from (1·12) $HkN(z) = \cos z$. From Massey & Bates (1943) $k = 4{\cdot}5 \times 10^{-18}$. Suppose that the temperature of the atmosphere is 1000°, then $H = 52{\cdot}6$ km. for oxygen, and taking $\cos z = \frac{1}{2}$ to represent average conditions then for the number of oxygen atoms per c.c. at the height where the maximum electron production occurs

$$N = 2 \times 10^{10},$$

which is the same order of magnitude as the number of particles per c.c. in the F region as determined by observation (Martyn & Pulley 1936). Supposing that the reaction (6·1) actually forms either the F_1 or the F_2 region, one cannot use the Pannekoek dilute equilibrium formula to get the electron density, because the radiative recombination coefficient corresponding to (6·1) is of order 10^{-12} (Massey & Bates 1943), while the observed value of α is of order 10^{-8} for the F_1 region and 10^{-10} or 10^{-11} for the F_2 region (§ 3). Since there are few negative ions per electron at such low densities (§ 2) then $n_e = n^+$ and $\alpha' = \alpha$ so far as reactions of type $O^- \rightarrow O + e$ are concerned. Massey & Bates (1943) attribute the observed high recombination to the reaction

$$O^+ + O^- \rightarrow O' + O'',$$

in which a positive ion combines with a negative ion to form two excited oxygen atoms. This reaction has so high a probability that it is important despite the comparatively low negative-ion population.

Let the number of transitions occurring per c.c. be defined as follows:

$$\begin{array}{llll} O^+ + O^- \rightarrow O' + O'' & \alpha'' n^+ n^-, & O' + O'' \rightarrow O^+ + O^- & \lambda n' n'', \\ O^- + h\nu \rightarrow O + e & n^- q^-, & O + e \rightarrow O^- + h\nu & \alpha^- n n_e, \\ O + h\nu \rightarrow O^+ + e & nq, & O^+ + e \rightarrow O + h\nu & \alpha' n^+ n_e. \end{array}$$

Then
$$\frac{dn^-}{dt} = \alpha^- n n_e - \alpha'' n^+ n^- - n^- q^- + \lambda n' n'',$$

$$\frac{dn_e}{dt} = -\alpha^- n n_e + nq + n^- q^- - \alpha' n^+ n_e.$$

Adding
$$\frac{d}{dt}(n^- + n_e) = nq + \lambda n' n'' - \alpha' n_e n^+ - \alpha'' n^+ n^-,$$

and in secular equilibrium $n^- + n_e$ is constant, so that

$$nq + \lambda n' n'' = n^+(\alpha' n_e + \alpha'' n^-).$$

If the atmosphere is electrically neutral, $n^+ = n_e + n^-$, and if $n_e \gg n^-$, then $n^+ = n_e$ nearly. Thus

$$nq + \lambda n' n'' = \alpha' n_e^2 + \alpha'' n_e n^-. \tag{6·2}$$

In full equilibrium the two processes balance independently, that is, $nq = \alpha' n_e^2$ and $\lambda n' n'' = \alpha'' n_e n^-$. In dilute temperature equilibrium, q will be multiplied by a dilution factor $\omega/8\pi$. If both populations of the excited states n' and n'' are of order ω in comparison with their full equilibrium values, the product $n'n''$ is of order ω^2, and $\lambda n' n''$ can be neglected in comparison with nq, giving

$$Q = \alpha n_e^2, \tag{6·3}$$

where α is now the effective recombination coefficient, equal to $\alpha' + \frac{n^-}{n_e}\alpha''$: its value may be taken to be 10^{-11} from observation.

If, however, one of the excited states is metastable, one population n' or n'' is not of order ω; and the product $n'n''$ is of the same order as q (the first order in ω), and since $\alpha'' n_e n^-$ is much greater than $\alpha' n_e^2$, $\lambda n' n''$ must be correspondingly greater than nq. The electron density will then be given by

$$\lambda n' n'' = \alpha n_e^2. \tag{6·4}$$

To investigate this point further the populations of metastable states in dilute temperature equilibrium must be examined.

7. Consider the populations of a number of states of an atom in secular equilibrium. Let n_i be the number of atoms per c.c. in the ith state and let $n_i A_{ij}$ be the number of transitions per c.c. per sec. from state i to state j. In secular equilibrium among m states

$$\sum_i n_i A_{ij} = \sum_j n_j A_{ji}, \tag{7·11}$$

both sums being from 1 to m with $i = j$ excluded. Defining A_{ii} by

$$A_{ii} + \sum_j A_{ij} = 0$$

with $j = i$ excluded from the sum, equation (7·11) becomes

$$\sum_{i=1}^{m} n_i A_{ij} = 0. \tag{7·12}$$

Let there be $(n_i)_T$ atoms per c.c. in state i in thermodynamical equilibrium at some temperature T (for a given number of atoms per c.c. in all states), and let $n_i = (n_i)_T \times t_i$. Write $(n_i)_T A_{ij} = (ij)$. Then $n_i A_{ij} = t_i(ij)$ and equation (7·12) becomes

$$\sum_{i=1}^{m} t_i(ij) = 0. \tag{7·13}$$

There are m such equations and $m-1$ unknowns (the ratios t_2/t_1 etc.), so that the determinant

$$\begin{vmatrix} (11) & (21) & (m1) \\ (12) & (22) & (m2) \\ (1m) & (2m) & (mm) \end{vmatrix}$$

must vanish and the quantities t_1, t_2, etc., are proportional to the minors of any row in the determinant.

In thermodynamical equilibrium $(ij) = (ji)$ and $t_1 = t_2 = \ldots = t_m = 1$. Under any other circumstances, if $E_j > E_i$, (ji) is an atomic constant (for bound states) and (ij) is proportional to the intensity of the radiation promoting the transition. In dilute radiation

$$(ij) = \frac{\omega}{8\pi}(ji),$$

and in a dilute radiation field which is weakened by absorption

$$(ij) = r_{ij}\frac{\omega}{8\pi}(ji).$$

To investigate the population of a metastable state at least three atomic states must be considered: the ground state $i = 1$, the metastable state $i = 2$ and a third state $i = 3$. Then

$$\frac{t_2}{t_1} = \frac{(12).(31)+(12).(32)+(13).(32)}{(21).(31)+(21).(32)+(31).(23)} \tag{7·2}$$

$$= \frac{A\omega' r_{12} + B\omega' r_{13}}{A + B\omega' r_{23}}, \tag{7·21}$$

where $\omega' = \omega/8\pi$, $A = (21).(31)+(21).(32)$ and $B = (31).(32)$. If the transition (21) is a normal one, t_2/t_1 is of order ω', but if state 2 is metastable and A is small compared with $B\omega'$,

$$\frac{t_2}{t_1} = \frac{r_{13}}{r_{23}}. \tag{7·22}$$

Thus the population of a metastable state is unity, not ω', at the top of the atmosphere (compared with its population at the solar temperature): but the metastable atoms are not distributed exponentially with height. That the population of a metastable state is of order unity in dilute equilibrium was first pointed out by Rosseland (1926).

8. Bates & Massey (1943) list eleven possible pairs of excited states O′ and O″. Of these six may show a large coefficient α''. They are

$$^1\mathrm{S}+{}^1\mathrm{S},\quad {}^3\mathrm{P}+{}^5\mathrm{S}^0,\quad {}^3\mathrm{P}+{}^3\mathrm{S}^0;\qquad {}^3\mathrm{P}+{}^5\mathrm{P},\quad {}^3\mathrm{P}+{}^3\mathrm{P},\quad {}^1\mathrm{D}+{}^5\mathrm{S}^0.$$

Of these the last three are ruled out if the value of the detachment potential is 3·0 V, as given by Vier & Mayer (1944). The states ^{1}S and ^{1}D are metastable. It is uncertain which pair of states gives the highest value of α'' (and λ), but since in one pair both the states are metastable, the product $\lambda n'n''$ must be greatest for that pair, namely,

$$\mathrm{O}^+ + \mathrm{O}^- \rightarrow \mathrm{O}(^1\mathrm{S}) + \mathrm{O}(^1\mathrm{S}).$$

Recognizing that more than one pair of excited states is possible one may write (6·4)

$$\Sigma\lambda_{rs} n_r n_s = \alpha n_e^2. \tag{8·1}$$

For each pair in full equilibrium

$$\lambda_{rs}(n_r n_s)_T = \alpha''_{rs}(n_e^2)_T,$$

and writing

$$\lambda_{rs}(n_r n_s)_T = \gamma_{rs}\Sigma\lambda_{rs}(n_r n_s)_T$$

(8·1) becomes

$$n_e^2/(n_e^2)_T = \Sigma\gamma_{rs} t_r t_s. \tag{8·2}$$

The quantities γ_{rs} are not, strictly speaking, independent of T, as the cross-sections for mutual neutralization of O$^+$ and O$^-$ (the quantities α''_{rs}) are not independent of T. In writing an equilibrium value of n_e one supposes that n is given.

Despite its approximate character equation (8·2) exhibits the important points of the argument. If $t_r = t_s = 1$, and if the r, s pair exhibits the largest coefficient so that γ_{rs} is nearly equal to unity, n_e will be equal to the number of electrons per c.c. in equilibrium with n oxygen atoms per c.c. *at the solar temperature* without any dilution factor. While γ_{rs} may have a high value, even approaching unity, it is doubtful whether t_r and t_s attain high values on account of absorption (the term r_{13} in (7·22)) and because in this case (the ^{1}S state of O) the condition $A \gg B\omega'$ is not satisfied. If t_r and t_s are very small equation (6·2) reduces to (6·3) rather than (6·4). The result may be expressed in terms of the solar temperature required to produce the observed electron concentration. Making no allowance for day and night averaging, and taking $n_e = 10^6$ and $n = 2\times10^{10}$, then (by day):

$\Sigma\gamma_{rs}t_rt_s$	$1·7\times10^{-3}$	4×10^{-4}	10^{-4}	0
Solar temperature	5000°	5500°	6000°	6500°

The entry 6500° for $\Sigma\gamma_{rs}t_r t_s$ zero assumes $\alpha = 10^{-11}$: other entries are unaffected by the choice of α. If $\alpha = 10^{-10}$ and the metastable states make no appreciable contribution, a temperature of 7100° is required.

If the solar atmosphere is in radiative equilibrium one would expect a lowering of the visible solar curve in the far ultra-violet, especially beyond 912 A, where one would expect very strong Lyman absorption. In the visible region radiation is received from depths considerably below, and therefore considerably hotter than, the extreme surface, which is at 4830°. An effective temperature in the Lyman continuum not much greater than 5000° might be expected. The populations of the metastable states of O are supported by radiation on the long-wave side of 912 A, so that the process $O(^1S)+O(^1S)\rightarrow O^+ + O^-$ is favoured relative to $O+h\nu\rightarrow O^+ + e$ if the solar-energy curve exhibits an absorption continuum at 912 A.

In any event it appears possible to account for the observed electron density in the F_2 region by supposing that the electrons are supplied by the ionization of O to O^+, either directly or by the double process $O(^1S)+O(^1S)\rightarrow O^+ + O^-$ followed by $O^-\rightarrow O+e$. With $\alpha = 10^{-11}$ the solar-energy curve required does not exceed the observed black-body curve at about 6800° in the visible spectrum, and a solar curve weakened below this will produce the observed ionization if the metastable states are abundant.

9. If t remained constant (not necessarily unity) throughout the atmosphere, a metastable state would yield a higher value of Q than the ground state of the same atom, because S is greater for the metastable state. In general the absorption coefficient per gram of the atmosphere is much lower for absorption by a metastable state than for absorption by the ground state. The layer formed by the metastable state would therefore be below the layer formed by the ground state and completely screen it from observation. With the ground state distributed exponentially with height and with t constant, in fact, only one layer would be seen, formed by that metastable state which gave the highest value of S. This does not, however, occur in practice because t is not constant as it contains the absorption term r_{13}. If r_{13} is sufficiently small, the metastable state will not form an observable layer because the population of the metastable state has not reached sufficiently large values to form a maximum electron density above the region where collisions become important. Below, say, 100 km. or so, collisions become more important than radiative processes in determining the populations of the metastable states. They will therefore be appropriate to the temperature of the earth's atmosphere and will not yield the full value of Q given by S/eH.

The two reactions

$$O(^1D_2)+h\nu\,(11{\cdot}6\,\mathrm{V})\rightarrow O^+ + e,\quad N(^2D^0)+h\nu\,(12{\cdot}1\,\mathrm{V})\rightarrow N^+ + e,$$

fall into this class. When the centre of the line $1\rightarrow 3$ is blacked out, r_{13} is proportional to $n_1^{-\frac{1}{2}}$, so that n_{12} is proportional to $n_1^{\frac{1}{2}}$ and still distributed exponentially. But the value of n_2 is still too small to produce an observable layer when $n_1 = 10^{13}$, and higher values of n_1 probably do not exist as collisions inhibit the dissociation of molecules into atoms.

Metastable molecular nitrogen offers possibilities. In a homonuclear molecule, vibration bands are forbidden by the $g \rightleftharpoons u$ rule, so that all the vibrational states of the lowest electron state $x\,{}^1\Sigma_g^+$ cannot perform a spontaneous transition to the ground state $x\,{}^1\Sigma_g^+$, $v'' = 0$. For states with $v'' > 6$ the energy necessary to ionize the molecule is less than 13·5 V, so that ionization can be accomplished by radiation which passes the barrier

$$O(^3P) + h\nu\,(13{\cdot}5\,V) \rightarrow O^+ + e.$$

If $v'' > 16$ the energy of the $x\,{}^1\Sigma_g^+$ state is greater than that of a ${}^1\Pi_u$ state so that $x\,{}^1\Sigma_g^+$, $v'' > 16$ is not metastable. An observable layer may be formed by

$$N_2(x\,{}^1\Sigma_g^+, 16 > v'' > 6) + h\nu\,(< 13{\cdot}5\,V) \rightarrow N_2^+ + e. \tag{9·1}$$

Too little is known about oscillator strengths concerned to subject to a quantitative test the hypothesis that (9·1) is responsible for the F_1 layer.

According to Mulliken (1932) molecular nitrogen should exhibit a ${}^1\Pi_g$ state, which would be metastable if it lay below ${}^1\Pi_u$. The ${}^1\Pi_g$ state has not been observed. The strongest unidentified band in the night-sky spectrum occurs at 3556 A. If this is the forbidden transition, ${}^1\Pi_g \rightarrow {}^1\Sigma_g$, the state has an excitation potential of 3·5 V and requires 12 V to become ionized. Such a metastable state would be excellently placed for forming the F_1 layer.

One may at this point discuss briefly another source of the electrons in the F_1 layer, namely, NO. The first ionization potential is 9·1 V, but the ionization continuum has not been observed (Price 1943). Since, however, examination of the N_2 absorption spectrum with high dispersion by Worley & Jenkins (1938) has revealed an ionization continuum at the first ionization potential which was at one time thought not to exist, or to be very weak (Price & Collins 1935) it may be supposed that higher dispersion may reveal a continuum for NO. A single band in the spectrum of NO^+ has been observed by Duffieux & Grillet (1936, 1937). The band head is at 5999 A. If this band is a transition to the ground state, the second ionization potential of NO is at 11·2 V.

Like every other constituent, NO would be distributed exponentially with height in an isothermal atmosphere in which there was no change in the intensity of photo-dissociating radiation. The population of NO in the actual atmosphere will accordingly be on an exponential curve at great heights at which both O_2 and N_2 are highly dissociated. The population of NO will fall off from this curve at just that point where absorption of the radiation which dissociates N_2 begins to be appreciable. This may be expected to occur somewhere near the height of the F_1 region.

10. Suppose that the electrons which form the F_1 region are supplied by absorption of light with energy below 13·5 V, the absorbing agent not being distributed exponentially with height. It is possible that several of the processes mentioned in § 9 contribute to the electron density. Radiation of energy 12·5 V will be diminished

in intensity, but the diminution need not be so great as to exclude the straightforward reaction

$$O_2(x^3\Sigma_g,\ v'' = 0) + h\nu\ (12{\cdot}5\,\mathrm{V}) \rightarrow O_2^+ + e$$

from forming the *E* region. As in N_2, vibration bands are forbidden and a number of vibrational levels are metastable. In addition the state $^1\Sigma_g$ is metastable, according to Mulliken. There is no doubt that O_2 can supply enough electrons to form the *E* region: the only difficulty is to explain why it does not supply so many that the *E* region will screen the upper layers. But the *E* region is at a considerably higher pressure than the upper regions, and it may be supposed that collisions are no longer negligible at this level, but prevent the metastable states of O_2 from building up too high a population.

The writer acknowledges with pleasure the assistance he has derived from discussing ionospheric problems with Dr D. F. Martyn and Dr C. W. Allen.

REFERENCES

Allen 1946 *Terr. Mag.* **51**, 1.
Appleton 1937 *Proc. Roy. Soc.* A, **162**, 451.
Bates & Massey 1943 *Phil. Trans.* A, **239**, 269.
Bhar 1938 *Indian J. Phys.* **12**, 363.
Chapman 1931 *Proc. Roy. Soc.* A, **132**, 353.
Duffieux & Grillet 1936 *C.R. Acad. Sci., Paris*, **202**, 937.
Duffieux & Grillet 1937 *C.R. Acad. Sci., Paris*, **205**, 39.
Martyn 1945 *Nature*, **155**, 363.
Martyn & Pulley 1936 *Proc. Roy. Soc.* A, **154**, 453.
Massey & Bates 1943 *Rep. Progr. Phys.* **9**, 62.
Mulliken 1932 *Rev. Mod. Phys.* **4**, 53.
Pannekoek 1926 *Proc. Acad. Sci. Amst.* **29**, 1165.
Price 1943 *Rep. Progr. Phys.* **9**, 13.
Price & Collins 1935 *Phys. Rev.* **48**, 714.
Rosseland 1926 *Astrophys. J.* **63**, 218.
Saha 1935 *Proc. Nat. Inst. Sci. India*, **1**, 222.
Saha 1937 *Proc. Roy. Soc.* A, **160**, 160.
Saha 1945 *Proc. Phys. Soc.* **57**, 271.
Vier & Mayer 1944 *J. Chem. Phys.* **12**, 28.
Woltjer 1925 *Physica*, **5**, 406.
Worley & Jenkins 1938 *Phys. Rev.* **54**, 304.

The statistical aspect of fatigue of materials

By A. M. Freudenthal, D.Sc., *Hebrew Institute of Technology, Haifa, Palestine*

(*Communicated by H. J. Gough, F.R.S.—Received* 10 *January* 1946)

The phenomenon which is usually described as 'fatigue' of materials, but for which the term 'progressive failure' would be more adequate, is the expression, on a macroscopic scale, of the progressive destruction of the cohesive bonds as a result of the repetitive action of an external load. It has the typical features of a mass phenomenon; both the cohesive bonds and the load repetitions are collectives in a statistical sense (Mises 1931). By applying the fundamental rules of the theory of probability many of the experimentally established relations between the principal variables can be theoretically deduced from the purely formal assumption of the existence of a statistical distribution function of the separation-strength of cohesive bonds.

1. Introduction

The very extensive experimental research concerning fatigue of materials in general and of metals in particular has resulted in the establishment of certain relations between the principal variables, such as amplitude, frequency and number of load cycles sustained, type of stress, size of sample and of crystalline grains, temperature, surface conditions and environment (Batelle Memorial Institute 1941). Attempts have not been lacking to formulate theories of fatigue, on the strength of which the principal facts and relations could be rationally explained (Gough 1926, 1931, pp. 207–227). However, the technique of fatigue testing is rather complex, and it is difficult to secure representative and reproducible results. Therefore, both the observations and the theories based thereupon have, to a certain extent, remained controversial.

Gough (1933, 1937), by accomplishing a synthesis of the accumulated evidence, has succeeded in presenting a rounded-off descriptive theory of fatigue. However, this theory, like all previous theories, falls short of the explanation of the essential problem of how the fatigue crack develops from the submicroscopic into the microscopic stage. The available experimental evidence is concerned with the propagation of the crack, once it has been formed and has attained microscopic size, not with its initial formation. No working hypothesis has, so far, been proposed concerning the submicroscopic happenings which, necessarily, precede the appearance of the visible cracks, although Gough has recognized the statistical character of the process (Gough 1937).

More recently Orowan (1939) presented a quantitative theory of fatigue which considers local plastic slip and subsequent strain-hardening as the principal cause of fatigue. He assumes that the 'total plastic strain', which is understood to be the sum of the absolute amounts of all positive and negative plastic strains, has a characteristic finite value. Since, for constant amplitude of the load cycle, this 'total strain' is a function of the sum of the areas enclosed by the hysteresis loops (damping), the above assumption very much resembles that of Ljunberg (1931), who contends that failure in fatigue is the effect of, or conditional upon the sum of the areas under the hysteresis loops of all load cycles sustained to reach a characteristic and finite value. However, this contention has never been supported by satisfactory evidence; on the contrary, the available evidence (Foeppl 1936) indicates that, under repeated loading, materials are generally able to dissipate an infinite amount of strain energy without fracture. It is true that the results of X-ray investigations support the assumption that the lattice destruction caused by plastic distortion tends towards a limit (Gough & Wood 1936), but this limit is not necessarily a function of the fatigue strength, nor is the progress of the distortion necessarily a measure of crack formation, particularly in the initial stages. If this were not so, the recorded amount of damping, being an expression of plastic distortion, would be a direct measure of progressive failure; however, such correlation has not so far been established (Brophy 1936). More recent reports (Koerber 1939, p. 26) even suggest that no extensive lattice destruction whatever occurs near the fatigue limit, but only accompanies the large distortions preceding single-stroke rupture. Furthermore, the assumption that fatigue is an effect of plastic slip and strain hardening is contradicted by the observation of a spreading fatigue crack in a perfectly brittle antimony crystal (Gough 1931, p. 215), as well as by the behaviour of quartz oscillators under large amplitudes.

It appears certain, therefore, that progressive failure is not inseparably associated with plastic slip but that the real fatigue effect is a large-scale expression of the progressive destruction of cohesive bonds (Gough 1926, 1931, pp. 207–227; Laute & Sachs 1929; Davidenkow & Schewandin 1931). Plastic slip and strain hardening are essential features only as far as they modify both the intensity and rate of bond destruction.

2. THE STATISTICAL APPROACH

It is assumed that long before any microscopic crack can be observed in a specimen subject to repeated load cycles, fatigue has already been initiated by the severance of a number of cohesive bonds. Moore (Batelle Memorial Institute 1941, p. 90) reports that in salvaging cracked railway axles by turning them down to the next smaller size it was necessary to remove $\frac{1}{16}$ in. below the apparent bottom of the crack. A specimen cut from the material which did not include the crack but included this $\frac{1}{16}$ in. had a lower fatigue limit than the virgin material while, if the $\frac{1}{16}$ in. was excluded the fatigue limit became normal. This observation suggests that the apparently sound material below the crack had already been damaged. Since such damage is neither revealed in metallographic examination nor through hardness tests or the examination of slip lines, and since it is not possible to trace all the local submicroscopic breakdowns which probably occur by any other means, such as X-rays, the only approach is to consider fatigue as an essentially statistical effect of happenings on a submicroscopic scale.

It will be shown that the purely formal assumption of the existence of a statistical distribution function of the separation strength of cohesive bonds within a quasi-homogeneous material is sufficient to explain the principal features of the fatigue phenomenon and to deduce the relations obtaining, at least in the initial stage. This assumption appears justified because, on the one hand, it is known that even under small loads which apparently cause perfectly elastic strain, very high stress concentrations of varying intensity occur within a polycrystalline substance as a result of dislocations and of the statistical orientation of the crystalline grains (Koerber 1939, p. 24); on the other hand, it is known that not only does the mechanical strength of a single crystal very widely differ in different directions, but as a result of inherent flaws and weak spots the separation strength in one and the same direction is a highly statistical quality and depends essentially on crystal size and surface conditions, as demonstrated by the observation of very high, but considerably dispersed strength values of samples the size of which approaches the dimensions of monomolecular chains (Joffé 1935).

A cumulative distribution curve of bond strength r has been presented in figure 1; the ordinate P_1 at the abscissa r_1 denotes the probability of occurrence within the material of cohesive bonds of a strength $r \leqq r_1$. This curve is the integral curve of the frequency distribution curve $p = f(r)$ (figure 2), where p denotes the probability that the separation strength of the bond is between r and $(r+dr)$, since the probability P_1 that the bond strength has a value equal to or smaller than r_1 is

$$P_1 = \int_0^{r_1} p(r)\, dr = f_1(r).$$

As the bond strength cannot be less than zero the lower limit is zero instead of $-\infty$. If an external load S is applied to the material, causing a macroscopic homogeneous state of stress, the forces s actually induced in the statistically orientated individual bonds will be distributed statistically about their mean value s_0. In one and the same bond the induced force s will vary with every repetition of the external load S as a result of changes in the internal structure caused by the effects of load and temperature (permanent deformation, slip, crystal break-up, reorientation of crystalline grains, heat development, etc.).

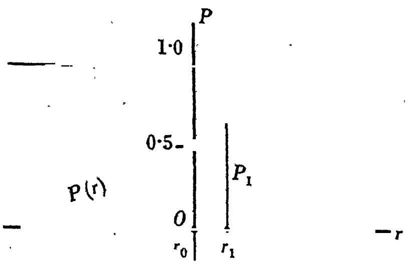


FIGURE 1. Cumulative frequency distribution curve of the strength of cohesive bonds.

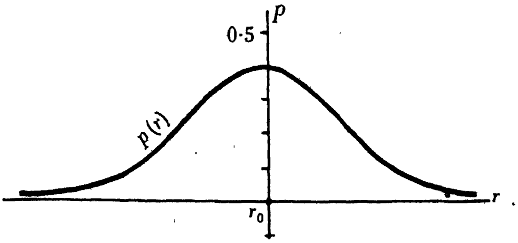


FIGURE 2. Relative frequency distribution of the strength of cohesive bonds.

The probability that a bond is destroyed as a result of the action of the external load S is the product of the individual probabilities of occurrence in the bond of forces s distributed about their mean value s_0 (which is a direct function of the force S), by the respective probabilities that the bond strength $r < s$. This combination of the frequency distribution $p = f(s)$ with the cumulative frequency curve $P_1 = f_1(r)$ results in a new cumulative distribution curve of the strength of cohesive bonds expressed in terms of the external load S.

In order to obtain this curve it must be considered that for every relation between the mean value s_0 of the forces s induced in the bonds and the bond strength r, the probability P of separation of a bond of strength r under the action of the load S which causes a distribution of forces s with the mean s_0, can be expressed as the sum of the individual probabilities of separation, that is as the sum of the products $f(s)f_1(r)$. Hence $P = \int_0^\infty f(s)f_1(r)\,ds$. By evaluating this integral for different magnitudes s_0 or S the ordinates of the new distribution curve $P = F_1(S)$ are obtained (figure 3). Each ordinate expresses the probability of the bond being destroyed if the external load attains the value of the abscissa S; it is, therefore, an expression of the bond strength in terms of the external load S. If all distributions are more or less normal (Gaussian), the shape of the curve $F_1(S)$ will rather closely resemble that of the cumulative distribution curve of bond strength $f_1(r)$.

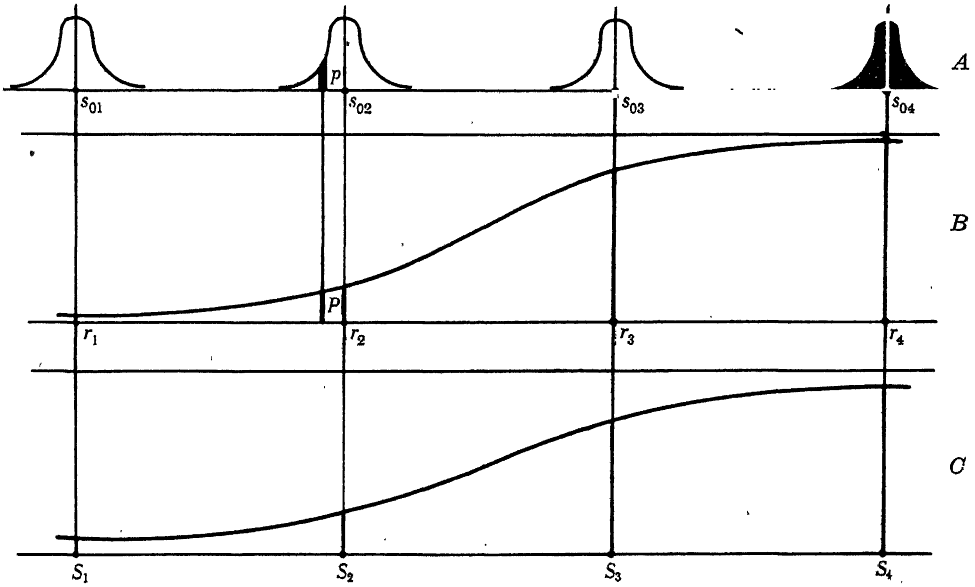


FIGURE 3. Development of the distribution curve of the strength of cohesive bonds in terms of the intensity of the external load, from the distribution of stress and strength in the cohesive bonds. A. Relative frequencies of forces in bonds $p = f(s)$ for various means s_0. B. Cumulative frequency of bond strength $P = f_1(r)$. C. Cumulative frequency of bond strength in terms of the external load $P = F_1(S)$, where $S = cs_0$. $S_n = \int_0^\infty f(s), f_1(r)\,ds$.

From the curve $P = F_1(S)$ the increase in the probability of rupture resulting from repetitions of the load S can be evaluated. If P is the probability that a cohesive bond will be destroyed by the action of the load S, $(1 - P)$ is the probability that it will not be destroyed. The probability that it will not have been destroyed after N repetitions of the load is, according to the rules of the theory of probability, $(1 - P)^N$. Consequently the probability that the bond will not sustain N repetitions without destruction becomes $[1 - (1 - P)^N]$, which is larger than P, and for constant P increases with N. Hence, the probability of rupture of the individual bond steadily increases with the number N of load cycles.

If, under a single-stroke load, the simultaneous severance of m bonds would result in a sudden total separation within the considered volume of the material, because the destruction of these m bonds would so considerably increase the forces in the remaining bonds that these would be rapidly destroyed, the probability that N load repetitions would *not* cause such separation since all m bonds remain undamaged is $(1 - P)^{mN}$; hence the probability of separation is $[1 - (1 - P)^{mN}]$. For a single-stroke load for which $N = 1$ this probability is $[1 - (1 - P)^m]$.

It has been assumed that non-rupture requires that m bonds remain undamaged under N load repetitions. Actually, the element under consideration will remain unbroken even if a certain number of bonds, smaller than m, across the potential area of separation is destroyed. The number of alternative states which may be defined as 'non-rupture' is, therefore, increased, with a consequent increase of the probability of non-rupture accompanied by a corresponding decrease in the probability of rupture. Each of the alternative states is characterized by a different probability P, since the destruction of even a small number of bonds increases the mean stress intensity s_0, and the probability of rupture of existing bonds becomes a function of the number of bonds already destroyed. The consideration of this fact would introduce considerable mathematical difficulties and prevent a simple statistical treatment of the problem; the probability of each alternative state would have to be computed separately since no statistical summation is possible. By carrying out this computation for theoretical samples consisting of a very small number of bonds and subject to a few load repetitions only, it can be shown that, under the assumption of a normal (Gaussian) distribution of individual bond strength, the numerical differences in the values of the probability of rupture obtained from the exact computation and from the expression derived from the assumption of non-rupture of all m bonds are relatively small, and do not materially affect the trend of the results.

Submicroscopic happenings are the cause of the *initiation* of fatigue cracks; the *propagation* of visible cracks is the final stage of fatigue, and it appears from observations that the test samples will have sustained some 50–90 % of the total number of load cycles before a crack becomes visible (Batelle Memorial Institute 1941, pp. 228–229) and large-scale happenings govern its further propagation.

3. Distribution curves of fatigue strength

From the cumulative curve $P = F_1(S)$ presented in figure 3 and assumed to be derived from a normal (Gaussian) frequency distribution, cumulative distribution curves $\Pi(S) = [1-(1-P)^{mN}]$ have been computed for values $mN = 10^1$ to 10^7. These curves, together with their first differentials, which are the respective frequency distribution curves $\frac{d}{dS}\Pi(S) = \pi(S)$ are shown in figure 4. It can be seen that with mN increasing the expected (most probable) strength, that is, the force S sustained without rupture during N load cycles which pertains to the highest probability of occurrence and which, therefore, is the mode of the frequency distribution, decreases; moreover, the dispersion of values expressed in terms of the standard deviation σ of the distribution $\pi(S)$ becomes narrower, and the negative skewness of the distribution (mode > mean) more prominent. This skewness is a characteristic feature of almost all observational series of mechanical strength.

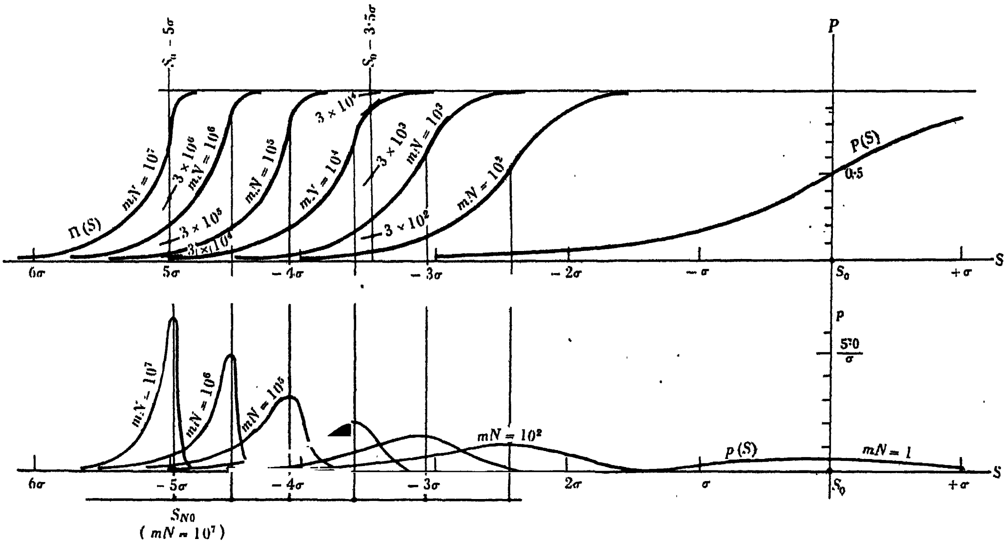


Figure 4. Cumulative distribution curves and frequency distributions of the strength of individual bonds and of groups of bonds under the action of repeated loads.

The mode of the frequency distribution for $N = 1$ pertaining to the cumulative curve $\Pi(S) = [1-(1-P)^m]$ indicates the expected (most probable) static strength.

Comparing the frequency distribution curves for the larger values of mN, and in particular the abscissae of their modes, the conclusion may be arrived at from figure 4 that the differences between the abscissae of the modes of consecutive curves are very nearly proportional to the differences of the logarithms of the respective values of mN, or

$$S_{N_0} - S_{n_0} = -k(\log mN - \log mn) = -k(\log N - \log n), \tag{1}$$

if S_{N_0} and S_{n_0} denote the most probable and therefore expected (modal) strength for N and n load cycles ($N > n$) respectively. For $n = 1$:

$$S_{n_0} = S_{T_0} \quad \text{and} \quad S_{N_0} = S_{T_0} - k \log N, \tag{1a}$$

the relation between the strength S_{N_0} under the action of N cycles and the static tensile strength S_{T_0}. Equations (1) or (1a) have the well-known form of the experimentally established relation $S(N)$ within the so-called unsafe range between the most probable or the average fatigue strength S_{N_0} and the number of load cycles N sustained; this is the range within which a finite number of load cycles only can be sustained without failure. That the relations observed express the effect of a general mechanism of fatigue, and are more than purely empirical lines of best fit of large numbers of observations, is suggested by the circumstance that they are valid for practically any kind of material, whatever its deformational performance, be it metal (Batelle Memorial Institute 1941, pp. 45–137), plastic (Lazan & Yorgiadis 1944; Batelle Memorial Institute 1941, p. 221) or textile fibre (Busse, Lessig, Loghborough & Larrick 1942). This fact may be considered an additional support of the assumption that fatigue is not associated with deformation, and that plastic slip and strain hardening, while constituting a prominent modifying effect, cannot be its underlying cause.

Equation (1) shows that the number m of bonds is irrelevant if samples of equal size are compared. For samples of different size and, consequently, of different number of bonds a definite size effect must be expected; as the probability of rupture under a definite load amplitude S increases with increasing number of bonds, the fatigue strength is bound to decrease with increasing sample size, a conclusion which is supported by experimental evidence (Batelle Memorial Institute 1941, p. 123).

The size effect in the stage of the initiation of fatigue cracks is modified by an opposing size effect which pertains to the stage of crack propagation: at equal rate of propagation the smaller cross-section will be destroyed more rapidly. The resulting effect depends, therefore, on the relative importance of the stages of crack initiation and of crack propagation in the process of rupture in fatigue.

The linear relation expressed in equation (1) is actually an approximation of a more complex relation; comparing the differences between the abscissae of the modes of consecutive frequency distributions $\pi(S)$ the value of k is seen to decrease slightly with increasing value of mN. The $S_{N_0} - \log N$ relation is, therefore, not a perfectly straight line but a curve slightly convex towards the N-axis; in the results of fatigue testing this convexity is sometimes more, sometimes less pronounced (Batelle Memorial Institute 1941, pp. 45–137).

In order to ascertain the variation in the range of dispersion of the fatigue strength, distributions $p(N)$ for the number N of load cycles of definite amplitude S_k sustained prior to rupture have been derived from the cumulative curves $\Pi(S)$ of figure 4. Their ordinates are direct functions of the inclinations of the curves $\Pi(S)$ at the abscissae S_k, and can be found by selecting out of the collectives $\Pi(S)$ of samples breaking at $N_1, N_2, \ldots, N_k$ load cycles respectively, the collectives of samples breaking at a definite load amplitude S_k or, actually, the group of samples breaking at a number of load amplitudes within a finite interval. Two such grouped distributions are shown in figure 5; their shapes indicate that the dispersion of the number of load cycles of constant amplitude leading to rupture increases with decreasing load amplitude. This conclusion is borne out by fatigue tests of a scope sufficient for statistical interpretation; it explains the fact that the results of fatigue tests are as a rule very widely dispersed.

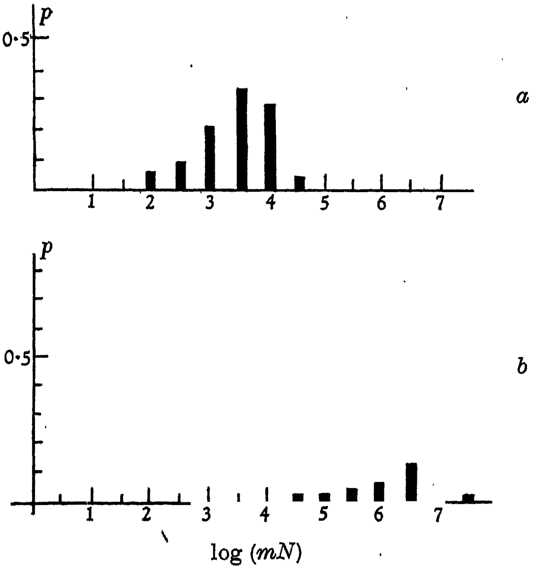


FIGURE 5. Frequency distribution of the number of load cycles sustained prior to rupture under load amplitude S_k. *a*. Load amplitude $S_k = S_0 - 3{\cdot}5\sigma$. *b*. Load amplitude $S_k = S_0 - 5{\cdot}0\sigma$.

$$p = \left[\frac{\dfrac{d}{ds}\pi(S)}{\Sigma\dfrac{d}{ds}\pi(S)}\right]_{S=S_k}$$

The rate of decrease with N of the expected fatigue strength is, according to figure 4, a direct function of the standard deviation σ of the initial distribution $P(S)$ and thus of its range of dispersion. The narrower this range the less rapid the decrease of strength as a result of load repetition. Every effect which tends to reduce the dispersion of the separation strength of the cohesive bonds will, therefore, increase the fatigue strength of the material. In metals, plastic slip followed by strain hardening is such an effect; the crystalline grains which initially are of various size, and are orientated at random, are gradually broken up to an approximately uniform size; this effect, together with a certain preferential reorientation both of grains and lattice, raises the uniformity of the strength of cohesive bonds. In dielectrics dipole orientation produces a similar result.

If plastic slip is not rapidly retarded and checked by strain hardening it results in the formation of transcrystalline microscopic cracks following the direction of slip. These cracks act as stress-raisers, since certain cohesive bonds will be brought into a state of severe strain.

The statistical fatigue mechanism illustrated in figure 4 is that of a brittle material. In metals showing plastic slip and strain hardening the progress of fatigue is retarded by the tendency of the initial cumulative distribution curve of the strength of cohesive bonds to become steeper as a result of the crystal break-up and rearrangement; the probability P of destruction of a bond at a certain load amplitude is gradually reduced to P' by strain hardening (figure 6); at the same time the probability is increased by the repetition of the load. The interplay of these two effects produces the essential behaviour of metals in fatigue. The more rapid and effective the break-up and reorientation of crystalline grains the more the destruction of bonds will be retarded. The capacity of a material for rapid rearrangement under relatively small loads appears, therefore, to be conducive to high fatigue performance. This suggests the existence of some correlation between fatigue performance and the shape of the relation between the damping capacity of a metal, as far as it is an expression of the extent of plastic slip and strain hardening, and the number of load cycles sustained.

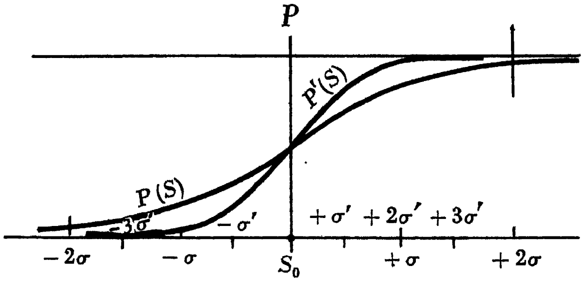


FIGURE 6. Schematic representation of the change of the cumulative distribution curve of bond strength as a result of plastic slip and strain hardening. $\sigma' = 0{\cdot}5\sigma$.

For metals this relation shows, according to Haigh (1928), three distinct stages (figure 7): the primary stage characterized by a brief evolution of heat and considerable but rapidly decreasing damping, accompanied by a substantial increase of hardness; the secondary stage during which some heat is still developed although both hardness and damping remain practically constant and the final, tertiary, stage immediately preceding rupture, during which damping steadily increases. The behaviour in the primary stage can be attributed to plastic slip retarded by strain hardening, that in the secondary stage to fatigue, that is to the progressive formation of submicroscopic cracks through severance of cohesive bonds, and that in the final stage to the propagation of microscopic and macroscopic cracks. Damping in the primary stage should be considered an expression of plastic slip and strain hardening not necessarily associated with fatigue; in the secondary stage it is a manifestation of the transformation of potential energy into thermal energy as a result of thermal vibration set up by the sudden separation of bonds and, to a small extent, into surface energy of the crack areas; in the final stage it is an indication of increased plastic slip due to the rapid increase of the number and extent of visible cracks which act as stress raisers. In this last stage damping will be a direct function of the number of cracks and discontinuities per unit volume which, necessarily, increases with increasing amplitude of load cycles.

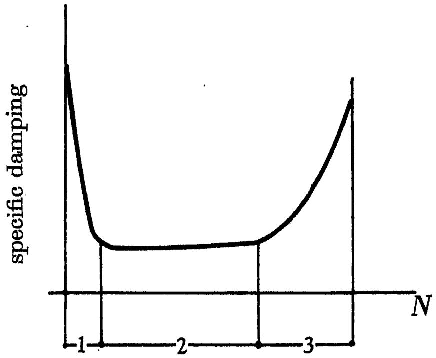


FIGURE 7. Relation between damping capacity and number of load cycles sustained for material showing plastic slip and strain hardening.

It is reasonable to expect that although the three stages are different in character, there will be no definite transition points but rather overlapping and superposition of consecutive stages.

It is probable that a material showing a conspicuous, but very short, primary stage (which would indicate a rapid and extensive internal reorientation resulting in a sharp reduction of the rate of progress of fatigue) and is accompanied by small damping during the secondary stage, will have a high fatigue performance. If, at a certain load amplitude, the reorientation is sufficient virtually to compensate the fatigue effect, this load amplitude delimits the safe range in which the probability of destruction of cohesive bonds has been reduced to such an infinitesimal value as to be practically nil even for millions of load repetitions.

At elevated temperatures no safe range can be observed for metals and equation (1) remains valid throughout (Gough 1931, p. 217). This behaviour is probably due to the elimination of strain hardening and crystal reorientation resulting from the effect of high temperatures. The relatively high fatigue strength at such temperatures might be an expression of the healing of destroyed bonds through increase of the amplitude of thermal vibrations.

4. DAMAGE THROUGH OVERSTRESS

If a load S_1 is repeated N_1 times and a load S_2 repeated N_2 times and the sample subsequently sustains N repetitions of the load S, and if the probabilities of rupture of cohesive bonds under the action of the loads S_2, S_1 and S are P_2, P_1 and P respectively, the probability of non-rupture of m bonds in (N_1+N_2+N) load cycles is $(1-P_1)^{mN_1}(1-P_2)^{mN_2}(1-P)^{mN}$; the probability of rupture is therefore

$$\Pi_1 = [1-(1-P_1)^{mN_1}(1-P_2)^{mN_2}(1-P)^{mN}]. \tag{2}$$

The probability of rupture pertaining to N cycles of the load S which is $\Pi = [1-(1-P)^{mN}]$ is increased by

$$(1-P)^{mN}[1-(1-P_1)^{mN_1}(1-P_2)^{mN_2}].$$

If $[1-(1-P)^{mN}]$ is considered a cumulative distribution curve of fatigue strength $\Pi(S)$, the probability of rupture at every load amplitude S repeated N times is increased by N_1 repetitions of the overload S_1 and the N_2 repetitions of S_2, and a new curve $\Pi_1(S)$ is obtained. The mode of the pertaining relative frequency distribution indicates the expected fatigue strength under N load cycles of a sample damaged previously by N_1 load cycles of amplitude S_1 and by N_2 load cycles of amplitude S_2; it is evidently lower than the expected fatigue strength for N load cycles only.

The progress of damage will, necessarily, be strongly affected by plastic slip and strain hardening. The more rapidly an internal reorientation proceeds, the more rapidly is the damaging effect of the overload reduced by the reduction of the probability of destruction of bonds. Metals of high performance under overstress will, therefore, be characterized by a shape of the relation between damping and number of load cycles whose primary stage is prominent but remains very short even under loads substantially exceeding the safe fatigue limit. On the other hand, a long-drawn-out primary stage of this relation is probably an indication of low resistance to damage by overstress, even if the amount of damping itself is considerable.

Understress, that is N_1 load cycles of amplitude S_1 below the fatigue limit S, remains practically without effect on the fatigue limit under N cycles of a brittle material if N_1 is smaller than N; this conclusion can be immediately deduced from figure 4. If the material shows plastic slip and strain hardening, the resulting internal orientation will even raise the initial fatigue limit, a conclusion which is supported by the available evidence.

5. NOTCH EFFECT

It has been assumed so far that an external load, producing a state of stress which, from the macroscopic point of view, can be considered as homogeneous, induces a statistically uniform distribution of the probability of destruction of cohesive bonds. Under an inhomogeneous state of stress, which may be caused by bending moments, notch effects, cracks or inhomogenetics inherent in the material, the probability of destruction of bonds becomes also a function of the geometrical co-ordinates.

If the stress distribution produced by a sharp notch is approximated by a discontinuous step function with two intensities, it can be assumed that a percentage μ of the m bonds is subject to the high stress intensity in the vicinity of the notch while the majority of bonds is in the field of the comparatively low homogeneous stress. If we denote by P the probability of destruction of a bond in the field of low stress intensity and by P_1 this probability in the immediate vicinity of the stress concentration, the probability that all m bonds remain unbroken under N load repetitions becomes

$$(1-\Pi_2) = [(1-P)^{(1-\mu)m}(1-P_1)^{\mu m}]^N. \tag{3}$$

The pertaining probability of rupture is therefore

$$\Pi_2 = 1-[(1-P)^{(1-\mu)m}(1-P_1)^{\mu m}]^N > \Pi, \tag{4}$$

where $\Pi = [1-(1-P)^{mN}]$, the probability of rupture without stress concentration.

From
$$(1-\Pi_2) = (1-\Pi)\left[\frac{1-P_1}{1-P}\right]^{\mu m N} \tag{5}$$

then
$$\log(1-\Pi_2) = \log(1-\Pi)-\mu m N\alpha, \tag{6}$$

where $\alpha = -\log\dfrac{1-P_1}{1-P} \sim (P_1-P)$ is a measure of the intensity of stress concentration.

Hence the probability Π_2 of rupture due to stress concentration in terms of that of the homogeneous field Π:

$$\Pi_2 = 1-(1-\Pi)\,e^{-\mu m N(P_1-P)}. \tag{7}$$

The probability increases rapidly with increasing value of $\mu m N\alpha$ (figure 8), that is, with increasing number of stress-raising effects μm, with increasing number of load repetitions and with increasing intensity of stress concentration. According to equation (7) these influences are interchangeable: a similar increase in the probability of rupture will be obtained either by increasing the number of stress raisers or by increasing the number of load repetitions. The severity of the notch effect in fatigue increases, therefore, considerably with increasing number of load repetitions. Even low-intensity stress concentrations characterized by a small difference between the probabilities P_1 and P are sharply magnified by the repetition of the load cycle, with the result that even a relatively small number of stress raisers of moderate intensity will be sufficient substantially to reduce the fatigue strength of a structural part, if the load is repeated frequently enough. This would explain the severe fatigue effect of one or a small number of tiny scratches or cracks on the surface of a polished sample, or of corrosion pits. Since the probability of rupture increases rapidly with the number of bonds affected by stress concentrations, the great number of corrosion pits and cracks explains their severe effect in reducing the corrosion-fatigue strength of the material, in spite of the low notch effect of the individual pit. The observed influence of the frequency of load cycles on the corrosion-fatigue strength, as well as the observation that there is no safe range in corrosion fatigue (Batelle Memorial Institute 1941, pp. 161, 164) can be explained as a result of the fact that in the corrosion-fatigue test the number of corrosion pits increases with testing time; that time is evidently a function of frequency and number of load cycles. The damaging effect of the increase in the number of stress concentrations thus tends to compensate the beneficial effect of plastic slip and strain hardening.

With regard to stress concentrations plastic slip can have a beneficial but also a damaging effect, according to its extent. It has already been mentioned that extensive slip resulting from high stresses produces dislocations and cracks in the direction of the gliding surfaces; these act as stress raisers. As their number may be assumed to increase with every repetition of the load their effect could be estimated by introducing μ as a direct function of N. The relation obtained by substituting $\mu = cN$ into equation (7) indicates an intensification of the damaging effect of load repetitions, the probability of rupture increasing in relation to N^2 instead of N.

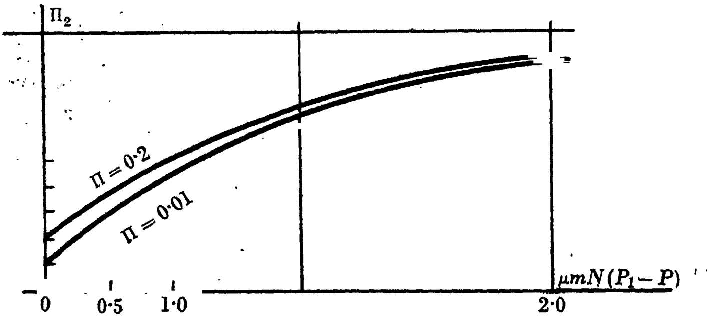


FIGURE 8. Increase of probability of rupture in fatigue as a result of stress concentration.

6. THEORY AND EXPERIMENT

Although the distribution function of the separation strength of individual cohesive bonds cannot be directly observed it is still possible to correlate the theoretical results with experiment and observation.

In order to obtain an undistorted picture of fatigue effects, plastic slip and strain hardening should be eliminated. This could be done by experimenting upon materials which, within the experimental range, are perfectly brittle or, when using technical metals, by either decreasing the testing temperature or increasing the testing speed sufficiently to ensure brittle behaviour.

For engineering purposes, however, it will not be necessary to eliminate the effects of plastic slip and strain hardening by creating special test conditions, since the distribution curves obtained under normal conditions, which represent the distribution of cohesive bond strength modified by plastic slip and strain hardening, represent the essential information required. In order that the pertinent curve of static strength should not be misleading as a result of the geometrical effect of local contraction which appears under normal test conditions, this test should be performed at the speed of the fatigue tests.

The cumulative frequency curve of static strength, which can be easily derived from an observed frequency distribution, represents the curve $\Pi = [1-(1-P)^m]$. From the recorded frequency distributions $p(N)$ of the number of load cycles sustained to rupture at various load amplitudes S the cumulative distribution curve $\Pi(S)$ can be derived for different values of N. The required collectives and their distribution $\Pi(S)$ are obtained by selecting from the observed frequency distributions $p(N)$ of the collectives of samples breaking at different load amplitudes $S_1, S_2, \ldots, S_k$ respectively; the collectives of samples breaking at those amplitudes at a definite number or within a definite range of numbers N of load cycles.

The curves $\Pi(S)$ are then fitted through the computed ordinates at the selected abscissae $S_1, S_2, \ldots, S_k$. The procedure is the reverse of that adopted in deriving the curves of figure 5 from those of figure 4. In the present case the former are observed and recorded and the latter derived therefrom.

The principal practical use which could be made of the curves $\Pi(S)$ so established is the evaluation of the damaging effect of various load amplitudes and numbers of repetitions, and of the effect of overstress by computing the appropriate curves $\Pi_1(S)$ and $\Pi_2(S)$ from the curves $\Pi(S)$. This is the main problem in the design of structures under repeated load cycles of varying amplitude; no other effective method is known so far, by which the fatigue effects of different load amplitudes and number of load cycles could be superimposed and the performance of structural materials under such conditions predicted on the strength of simple fatigue tests.

REFERENCES

Batelle Memorial Institute 1941 *Prevention of the failure of metals under repeated stress.* New York: Wiley and Sons.

Brophy, G. R. 1936 *Trans. Amer. Soc. Metals*, **24**, 154–185.

Busse, W. F., Lessig, E. T., Loghborough, D. L. & Larrick, L. 1942 *J. Appl. Phys.* **13**, 715–724.

Davidenkow, N. & Schewandin, E. 1931 *Metallwirtschaft*, **10**, 710–714.

Foeppl, O. 1936 *J. Iron Steel Inst.* (II), **134**, 393–455.

Gough, H. J. 1926 *The fatigue of metals.* London: Ernest Benn Ltd.

Gough, H. J. 1931 The present stage of knowledge of fatigue of metals. *Congr. Int. Ass. Test. Mat.* Group A, **1**.

Gough, H. J. 1933 *Proc. Amer. Soc. Test. Mat.* **33**, 3–101.

Gough, H. J. 1937 Characteristic of the deformation and fracture of metals as revealed by X-rays. *Congr. Int. Ass. Test. Mat.* Group A, Report no. 34.

Gough, H. J. & Wood, W. A. 1936 *Proc. Roy. Soc.* A, **154**, 510–539.

Haigh, B. P. 1928 *Trans. Faraday Soc.* (II), **24**, 125–137.

Joffé, A. 1935 The mechanism of brittle rupture. *Int. Conf. Phys.* **2**, 79.

Koerber, F. 1939 Das Verhalten Metallischer Werkstoffe im Bereich Kleiner Verformungen. *Proc. Fifth Int. Congr. Appl. Mech.* 20–33. New York: Wiley and Sons.

Laute, K. & Sachs, G. 1929 *Mit. Mat. PrüfAmt.* **9**, 89–91.

Lazan, B. J. & Yorgiadis, A. 1944 *Amer. Soc. Test. Mat.* **44**, 78.

Ljungberg, K. 1931 Konstante Brucharbeit als Erklärung für den Bruch durch Ermüdung. *Congr. Int. Ass. Test. Mat.* Group A, **1**, 337–338; 1930 preliminary report A, pp. 149–154.

Mises, R. v. 1931 *Probability and its applications*, p. 19. Leipzig.

Orowan, E. 1939 *Proc. Roy. Soc.* A, **171**, 79–105.

Kinetic friction in or near the boundary region

I. Apparatus and experimental methods

By B. Chalmers, P. G. Forrester and E. F. Phelps

(*Communicated by E. N. da C. Andrade, F.R.S.—*
Received 14 *February* 1946.—*Read* 9 *May* 1946)

[Plate 6]

When the hydrodynamic forces in a lubricant between two sliding surfaces are insufficient to carry the load between them, the condition which occurs is one of mixed boundary and fluid lubrication. From the point of view of damage to the surfaces and of resistance to sliding, the important part is the boundary friction, which can conveniently be studied only when isolated from fluid friction. There are two methods of obtaining relatively pure boundary friction: first, to ensure that the amount of lubricant present is only sufficient to provide a boundary film, or secondly, so to control conditions of load, speed and contact area that little or no opportunity is given for fluid film formation. The relative advantages of the two methods are discussed, and apparatus is described in which the second method is employed. In this apparatus one specimen is driven at a fixed velocity, and a second specimen is applied to it by a dead load. The force required to restrain the second specimen is measured by balancing against a variable dead load, instability being suppressed by fluid damping. A direct reading of coefficient of friction is thus obtained by a null method. Reproducibility of determinations is generally within 10 %, the limiting factor being the accuracy with which exactly similar surfaces can be reproduced.

Introduction

Kinetic friction between two solid surfaces is commonly divided into three categories:

(1) Dry friction between chemically clean surfaces.

(2) Fluid friction between surfaces separated by a lubricant film of such thickness that friction is determined solely by the hydrodynamic properties of the fluid.

(3) Boundary friction between surfaces separated by a lubricant film so thin that the hydrodynamic properties of the fluid play no part in determining the frictional behaviour.

This work is concerned with kinetic friction occurring between surfaces in the presence of a lubricant. In such circumstances relative motion of the surfaces brings about hydrodynamic pressure on the lubricant which supports a part of the load between the surfaces. This hydrodynamic pressure may be derived only from a departure from absolute parallelism of the surfaces which results in the formation of an oil wedge or wedges (apart from a small effect due to the inertia of the lubricant, referred to by Archbutt & Deeley (1912)). In a journal bearing such a wedge results from the difference between the diameters of the journal and the shaft, while in a Michell thrust bearing it is derived from the tilt of the pads. Between two 'plane' surfaces, the wedges are on a much smaller scale and result, as Karelitz (1940) points out, from the imperfections in the surfaces. In any of these circumstances the lubricant will exert a definite pressure at each point on the surfaces, tending to separate them. If the summation of all these separate forces multiplied by the area over which they act results in a force equal to the normal load between the surfaces, full fluid lubrication occurs. In these circumstances no damage to the surfaces can result from their relative motion, and the resistance to motion is provided entirely by the internal friction of the lubricant. If, however, the normal load between the surfaces exceeds the sum of the lubricant film forces, the surfaces approach one another until high spots are separated only by a boundary film of lubricant. In these circumstances damage to the surfaces may result, and resistance to sliding is derived not only from the internal friction of the lubricant, but also from the much greater boundary friction at the high spots. Thus, although the greater proportion of the contact area is still under fluid film conditions, it is the much smaller area, where such a film is absent, that is of interest from the point of view of damage to the surfaces and of resistance to sliding. Hence arises the importance of examining the frictional phenomena occurring under boundary conditions. The object of the present work was to design apparatus for examining such phenomena by measurement of the coefficient of friction under different conditions, and for comparing the boundary frictional properties of different combinations of surfaces and lubricant.

THE PRINCIPLES AND DESIGN OF THE APPARATUS

The first essentials of an apparatus to measure coefficient of friction are that surfaces of two materials shall be brought into contact by a known normal force, that a definite relative velocity be imparted to the surfaces, and that the tangential force between them be measurable. These requirements may most conveniently be met by driving one specimen and measuring the force required to restrain the other. The construction of an apparatus to fulfil these fundamental requirements and to constitute a satisfactory measuring instrument involves certain practical problems. These problems, and the manner in which they have been met, may be discussed under the following headings.

(1) *Obtaining an approximation to true boundary conditions*

As has been noted above, in almost all practical circumstances boundary friction occurs only locally between two surfaces, fluid film conditions existing over most of the apparent area of contact. In order to study boundary friction it is necessary to isolate this phenomenon as far as possible by using conditions different from those of practice. This may be done by either of two methods, or by a combination of the two:

(*a*) By so choosing the geometry of the surfaces that the real area of contact corresponds as closely as possible with the apparent area of contact (thus eliminating oil pockets), and by using very low sliding speeds. This technique has been used by a number of previous investigators (e.g. Bowden & Leben (1940)).

(*b*) By applying to the specimens a film of lubricant of a thickness, corresponding to a small number of molecules. This may be done (i) by the Langmuir and Blodgett technique described in Blodgett (1935) and used by Bowden & Leben (1940) and by Frewing (1942, 1943); (ii) by applying excess of lubricant and rubbing down with several changes of clean cloth (developed by Blodgett & Langmuir (1937), and by Germer & Storks (1939), and used by Dacus, Coleman & Roess (1944)); (iii) by deposition from a volatile solvent. In the presence of such a thin film of lubricant, fluid film lubrication cannot obtain, so it is possible (theoretically at least) to use any sliding velocity and still maintain boundary conditions.

Certain difficulties are associated with the second method. First, the lubricant film applied by any of these methods may have different properties from that which occurs in practice as a result of excess being squeezed out under pressure (experiments described later show that such a difference does exist). Secondly, there is uncertainty as to the number of molecular layers which constitute a boundary film. Bowden & Leben (1940) have shown that a unimolecular layer of stearic acid suffices to give the minimum friction obtainable with this lubricant, whereas with cholesterol more than nine molecular layers are required. Thirdly, a thin film is worn away during sliding, so when such a film is employed the same track must not be retraversed. This makes the study of running-in effects impossible, and, if tests at relatively high velocities are required, necessitates either a very rapid frictional determination or some experimental arrangement which allows for a long continuous run without the same track being retraversed.

The use of high speeds and thin lubricant films is not, therefore, a very practical proposition for lubricant testing. The technique of applying thin films has, however, as will be described later, been of considerable value in the present series of tests, when used in conjunction with tests with excess of lubricant.

The use of low speeds and restricted areas of contact has been adopted with the present apparatus. The usefulness of this method depends upon a knowledge of the relationship between friction and sliding velocity, in order that deductions of the behaviour at relatively high velocities may be made from observations at low velocities. In Part II of this paper experiments designed to examine this relationship will be described.

Of the various arrangements of specimen geometry by which a restricted area of contact may be obtained, that chosen was a hemisphere in contact with a flat plate. The plate was made the driven specimen, and the hemisphere the restrained specimen.

(2) *Movement of the driven specimen*

The driven specimen must move steadily at a velocity which is independent of the frictional force imposed on it by the restrained specimen, but which can be varied within wide limits. It was also considered desirable that the driven specimen should be capable of an indefinitely long continuous movement. This implies rotational rather than rectilinear motion. The method of drive adopted after a number of experiments was by a worm drive from an electric motor through a specially designed gearbox.

(3) *Mounting of the restrained specimens*

The restrained specimen must be so mounted that when measurements are made no significant force acts on it other than the frictional force and the restraint of the force-measuring system. The use of a single specimen involves introducing some means of counteracting the tilting force, so that it is difficult to comply with this condition. A better method is to employ three specimens placed at the corners of an equilateral triangle and fixed in a rigid holder. If the normal force is applied at the centre of this triangle, which is also made the centre of rotation of the driven specimen, then each specimen carries one-third of the normal load and, assuming the coefficient of friction to be the same in each case, the tilting and translational forces acting on the specimen-holder are balanced. The only resultant force is rotational. This arrangement of specimens is employed, with a slight modification, in that the sliders are placed at slightly different distances from the centre so that their tracks do not coincide. The radial force resulting from this is very small.

(4) *Measurement of frictional force*

It has been noted under (3) that at the time measurements are made it is essential that the only forces acting on the restrained specimen should be the frictional force and the restraint of the force-measuring system. Under these conditions there is a tendency to instability in certain circumstances. This has been demonstrated by Bowden & Leben (1939), by Papenhuyzen (1938) and by Morgan, Muskat & Reed (1941). The instruments used by these investigators were essentially similar in principle. In each case the applied load is maintained constant, one specimen driven at a fixed velocity while the other is restrained by a spring system, and the changes in displacement of the restrained specimen are recorded. With this type of apparatus three types of behaviour of the restrained specimen have been observed:

(*a*) It may take up a constant position, corresponding to a certain deflexion of the spring system and thus to a constant frictional force.

(*b*) It may move forward at a velocity equal, or nearly equal, to that of the driven specimen until a definite deflexion is reached, when it slides rapidly back to a position of smaller deflexion, then again moves forward and repeats the cycle.

(*c*) It may vibrate in an approximately sinusoidal manner.

Bowden & Leben believe that this behaviour is evidence that friction may, in some circumstances, be discontinuous, and associate it with the making and breaking of minute welds. Kaidanovsky & Haykin (1933) and also Blok (1939) have shown, however, that vibrations of either the 'relaxation' or the sinusoidal type may be explained by the assumption that, under the particular conditions in which they occur, the frictional force decreases with increasing velocity. Under these conditions the equilibrium which may theoretically be established when the frictional force is equal to the restraining force of the spring is unstable. For, consider the effect of a momentary slight increase in the frictional force. The relative velocity of the two specimens is thereby reduced, further increasing the frictional force, so that the relative velocity of the two specimens decreases until it reaches the value corresponding to a maximum frictional force. The specimen then moves forward until the new value of the frictional force is exceeded by the increasing restoring force of the spring. The specimen then commences to slip back, and this increases the relative velocity, reducing the frictional force. The slip, therefore, continues until the restraining force of the spring again becomes less than the frictional force, when the cycle is repeated. The system is, therefore, of an inherently vibratory character.

On the other hand, if conditions are such that the frictional force increases with increasing velocity, the system may be shown to be stable, so that the specimen maintains a constant deflexion.

For any form of oscillation to be performed by the restrained specimen, it is an essential condition that the force-measuring system is in the nature of a spring so that the restraining force depends upon the displacement of the specimen. This system is not used in the present apparatus as the principal method of force measurement, for it depends upon accurate calibration and accurate zero determination, both of which are difficult and must frequently be repeated. Instead, the method of balancing against a dead load is employed. The restraining force applied to the specimen is then independent of its displacement. Under these conditions instability will still manifest itself when friction decreases with velocity, for on any slight disturbance from equilibrium the specimen will move continuously forward or backward, according to the direction of that disturbance.

Some means must therefore be employed to impose a damping force on the specimens. This force must be zero when the specimens are stationary, but change more rapidly than the frictional force with changing velocity, in such a way as to compensate for the change of frictional force. Experiments showed that adequately heavy damping could be obtained by rigidly attaching to the specimen-holder pistons moving in cylinders full of a viscous liquid, the clearances being fairly small.

The dead load used to restrain the specimen has been made continuously variable by a water-siphoning system so that it may be adjusted to balance the frictional force, and so keep the restrained specimen stationary. This constitutes a null method of measuring coefficient of friction, in which the null point is represented by zero velocity of the restrained specimens. This method is entirely satisfactory when, as in most cases, the time required to take a reading is not an important factor. When rapid measurements are important, as when studying short-period variations in the coefficient, a small additional restraining force may be applied by a spring, the deflexion of which gives a measure of that force. Since this additional force represents only a fraction of the total restraining force, the possible inaccuracies of the spring method are proportionately reduced.

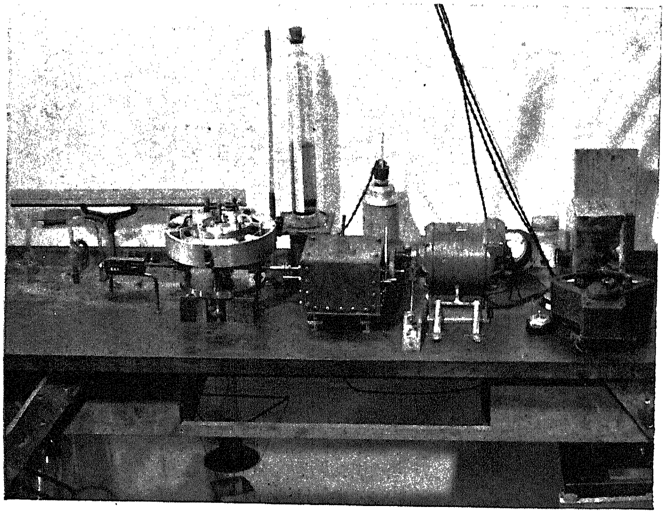


FIGURE 1. General photograph of apparatus.

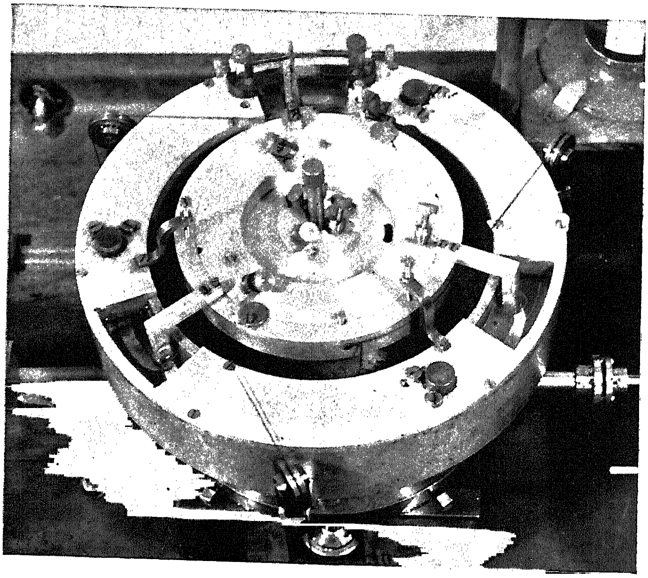


FIGURE 2. Photograph of friction-measuring instrument.

THE CONSTRUCTION OF THE APPARATUS

Figure 1, plate 6 shows the general lay-out of the apparatus, while figure 2, plate 6 is a photograph of the friction-measuring instrument.

A diagram of the instrument is shown in figure 3, and letters in the description below refer to those in this figure.

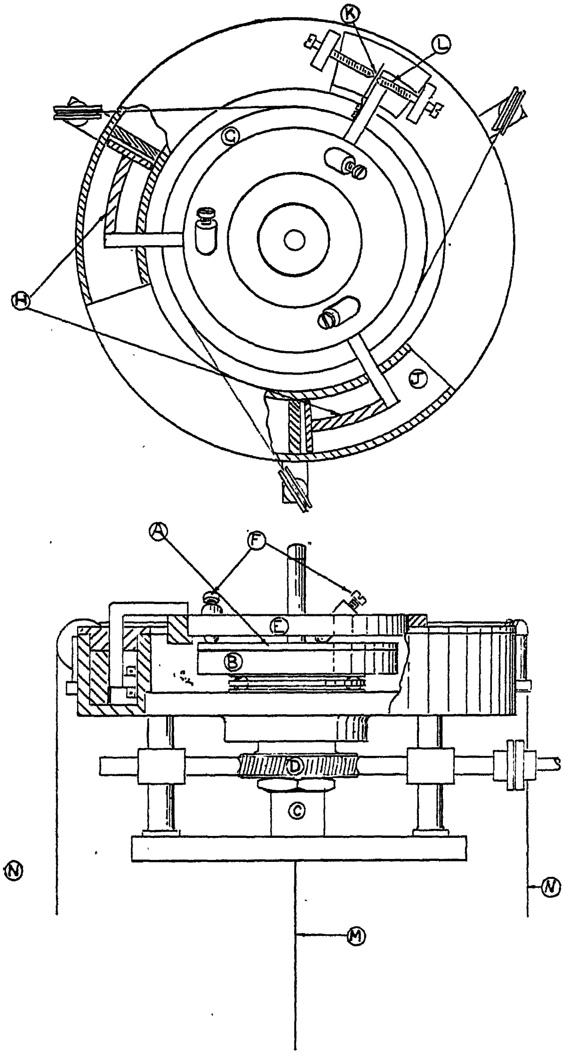


FIGURE 3. Diagram of instrument.

The driven specimen *A*, in the form of a flat annulus, is placed on a heavy gun-metal disk, *B*, and held in position by pegs. The disk, *B*, is mounted on ball races, and is attached to a hollow shaft, *C*, which is rotated by means of a worm-and-wheel mechanism, *D*. The shaft carrying the worm is rotated, through a five-speed gearbox, by an induction motor. The disk, *B*, is wound internally with nichrome wire, through which an electric current may be passed to heat the driven specimen.

The sliding specimens are in the form of hemispherically ended cylinders. These are mounted in a specimen-holder, *E*, which is a Dural disk machined to a thin section except at the rim and the centre. The rim contains three holes drilled at 45° to the vertical in a tangential plane and at intervals of 120° round the circumference. The three sliders are inserted in these holes, which are at slightly different distances from the centre. The amount by which the sliders project from the undersurface of the holder is adjusted by means of the screws, *F*. When correctly adjusted the sliders are locked in place by grub-screws. Underneath the holder, and exactly at its centre, is a steel ball which fits into a cylinder and thus keeps the holder central. To the bottom of this ball is attached a hook, from which a weight is suspended. The specimen-holder fits into an annulus, *G*, and is locked in place. To the annulus, *G*, three radial arms are attached; which are in turn attached to curved arms carrying pistons, *H*, which dip in an annular bath, *J*. The bath, *J*, has a liquid-tight cover, complete except for three gaps through which the piston arms enter. The three separate covered portions thus formed are closed at one end, forming curved 'cylinders' of rectangular section in which the pistons move. The bath is completely filled with a viscous liquid of low vapour pressure. Rotation of the specimen-holder can thus only take place by leakage of the liquid between the pistons and the cylinder wall. Thus very powerful damping is obtained. The clearance between the pistons and the cylinder wall is about 0·01 in. at the sides, and 0·025 in. top and bottom. The adjusting screws, *F*, compensate for different thicknesses of plates and different lengths of sliders, which would otherwise affect the top and bottom clearances. Pointers indicate the correct position of the holder.

The weight of the specimen-holder and the annulus in which it is placed is thus entirely supported by the sliders (with a small correction for the buoyancy of the pistons in the damping medium). Additional normal load is applied through a thin steel wire, *M*, attached to the ball under the specimen-holder, and passing through the centre of the apparatus.

Three small bolts are inserted into the outer rim of the annulus, arranged at the corners of an equilateral triangle. To these bolts are attached threads, *N*, which pass over pulleys to the corners of a very light wooden triangle, from which is suspended a light aluminium container. The triangle with the container and its contents form the friction-balancing load. The pulleys are mounted on ball races, and the frictional drag they introduce represents less than 2 % of the friction-balancing load. A glass tube, independently suspended, passes almost to the bottom of the container and is connected through a stopcock to a measuring cylinder mounted above the vessel. The only other outlet to the measuring cylinder is through a two-way stopcock, whereby the cylinder may be connected alternatively to a pressure or vacuum receiver. The pressure and vacuum receivers are connected to the high- and low-pressure sides of a reciprocating pump. Thus according to the way the two-way stopcock is turned, water may be forced from the container to the measuring cylinder or from the cylinder to the container. Thus the friction-balancing load is rapidly and continuously variable.

As mentioned above, there is a subsidiary method of applying a restraining force to the specimen-holder, utilizing a spring. A short steel spring, K, is attached radially to the rim of the annulus, G, with flat silver contacts on each side of its free end. Forward or backward movement of the specimen-holder brings this spring into contact with one of the silver-tipped bolts, L, fixed in brass blocks and insulated from the instrument. Contact of the free end of the spring with one of these bolts completes a circuit which lights a small lamp. When contact is established by movement of the specimen-holder, the friction-balancing force is reduced or increased by an amount which depends upon the deflexion of the spring.

Two separate lamp, mirror and scale systems are included in the apparatus. The first of these is for measuring the velocity of the driven specimen. A hexagonal collar is attached to the shaft, C, to each face of which is attached a mirror. One or other of these mirrors throws the image of a lamp on to a curved scale. The scale is calibrated in mm. movement of the table at the distance from the centre at which the specimens are placed. The velocity at any given gearbox setting can then be measured by noting the time for a given specimen movement. This lamp and scale system also enables any irregularity of movement of the driven specimen to be observed.

The second optical system is for measuring the movement of the specimen-holder. A mirror is attached above the specimen-holder and at its centre. This mirror throws the image of a lamp on to a plane mirror, which in turn throws the image on to a scale mounted above the apparatus. A long optical lever is thus obtained. This optical system serves two purposes. First, it acts as a sensitive means of determining the position of balance between frictional force and friction-balancing load, at which position there should be no resultant movement of the specimen-holder. Secondly, when the spring system is in use, it serves to measure the movement of the specimen-holder and therefore the deflexion of the spring. The zero is established by moving the specimen-holder until the lamp just lights. Calibration of the spring is carried out by repeated observations on the same system of the deflexions at different values of the friction-balancing load.

OPERATION OF THE INSTRUMENT

The specimens are placed in position and the height of the holder adjusted by the screws, F. The driven specimen is then set in rotation at a speed which depends on the setting of the gearbox. The friction-balancing load is then adjusted until no resultant movement of the specimen-holder takes place. In practice there is always some slight backward-and-forward movement of the holder due to minor fluctuations in the coefficient. The required friction-balancing load having been noted, the gearbox setting is changed, and a reading taken for each velocity. The velocities available are 0·0096, 0·0328, 0·122, 0·526 and 2·26 cm./sec.

Preliminary experiments with various materials showed that marked changes in the coefficient occur with repeated movement of the sliders over the same tracks. Accordingly, determinations of the coefficient at each velocity are made initially and after various numbers of completed revolutions. Since the specimens are at slightly different distances from the centre, each makes a track which is independent of that of the other two. A revolution-counter attached to the driving shaft serves to indicate the total number of revolutions.

TESTS ABOVE ROOM TEMPERATURE

For tests above room temperature, current is passed through the winding in the disk, *B*, enabling temperatures up to about 150° C to be obtained. No thermostatic control is used, but a simple device in series with the heater provides compensation for variations in the room temperature and the mains voltage. This is found to maintain the temperature within ± 1° C of that required. Temperature is measured by a thin-wire thermocouple in sliding contact with the upper surface of the specimen, at the same distance from its centre as the specimens.

PREPARATION OF THE SPECIMENS

Sliders are, in general, prepared by dry machining in a lathe to give a hemispherical end with no subsequent treatment. Since they are set at an angle to the driven specimen (the plate) a freshly cut surface is exposed to it.* Various different methods are used for preparing the plate, the one generally adopted being to grind with emery cloth on the face-plate of a lathe, swab with caustic soda solution till complete wetting is obtained, then to dry with acetone or isopropyl alcohol.

REPRODUCIBILITY OF RESULTS

A difference between the frictional force and the friction-balancing load of 4 g. is sufficient to bring about a readily observable movement of the specimen-holder. With a normal load of 2 kg. and a coefficient of friction of 0·1, this represents a sensitivity of 2 %. In general, successive determinations of the coefficient at a definite velocity made upon the same plate, sliders and lubricant without repreparation differ by less than this figure. If, however, a number of tests are made with the same materials but with the specimens reprepared by the same method for each experiment, results may differ by 10 % or, in a few cases, rather more. It is believed that this 'scatter' of results must be attributed to the difficulty of obtaining similar surface finishes with successive repreparations, even though finishing methods are standardized as closely as possible. This degree of reproducibility is sufficient for most purposes, and, of course, more accurately representative figures may be obtained by taking the means of a number of tests with a fresh preparation each time. In some experiments it is possible to use the same specimens, without repreparation, for a number of tests, for example when it is required to investigate the effect of temperature, or of thickness of the lubricant film. This method was widely used in a series of tests to be described in Part II.

---

* The tip of a specimen turned in a lathe to hemispherical shape is pushed aside rather than cut. Hence the advantage of setting specimens at an angle.

REFERENCES

Archbutt, L. & Deeley, R. M. 1912 *Lubrication and lubricants.* London: Charles Griffin.
Blodgett, K. B. 1935 *J. Amer. Chem. Soc.* **57**, 1007.
Blodgett, K. B. & Langmuir, I. 1937 *Phys. Rev.* **51**, 904.
Blok, H. 1939 *World Automotive Engng Congress*, May 1939.
Bowden, F. P. & Leben, L. 1939 *Proc. Roy. Soc.* A, **169**, 371.
Bowden, F. P. & Leben, L. 1940 *Phil. Trans.* A, **239**, 1–27.
Dacus, E. N., Coleman, F. & Roess, L. 1944 *J. Appl. Phys.* **15**, 813.
Frewing, J. J. 1942 *Proc. Roy. Soc.* A, **181**, 23.
Frewing, J. J. 1943 *Proc. Roy. Soc.* A, **182**, 270.
Germer, L. H. & Storks, A. H. 1939 *Phys. Rev.* **55**, 648.
Kaidanovsky, N. L. & Haykin, S. E. 1933 *Z. Tech. Phys.* **3**, 91.
Karelitz, C. B. 1940 *Proc. Special Summer Conferences on Friction and Surface Finish, Mass. Inst. Tech.*
Morgan, F., Muskat, M. & Reed, D. W. 1941 *J. Appl. Phys.* **12**, 742.
Papenhuyzen, P. J. 1938 *Ingenieur, Utrecht*, **53**, 75.

Kinetic friction in or near the boundary region

II. The influence of sliding velocity and other variables on kinetic friction in or near the boundary region

BY P. G. FORRESTER

(*Communicated by E. N. da C. Andrade, F.R.S.—Received* 14 *February* 1946.—*Read* 9 *May* 1946)

An investigation has been made of the causes underlying change of friction with sliding velocity. The method adopted was to measure the friction of several different combinations of materials under three different conditions: dry, with excess of various lubricants, and with thin films of various lubricants applied by two different methods. The experiments made show that changes in friction with velocity may be derived from at least three sources. First, such a change may be derived from the properties of the clean metal surfaces. Certain combinations of materials show, when unlubricated, a decrease in friction with increasing velocity, and this tendency may remain in a modified form even when a lubricant is added. Various hypotheses to explain this tendency are discussed. Secondly, with a boundary film of lubricant present, friction may increase with increasing velocity, and the evidence suggests that this change is due to partial destruction of the boundary film, the rate of which destruction rises with increasing sliding speed. Thirdly, when there is excess of lubricant, fluid film effects may occur even with restricted areas of contact and sliding speeds as low as 0·5 cm./sec.; such fluid film effects bring about a decrease in friction with increasing velocity. This tendency is most marked when one component is a soft metal, and may well explain the good 'antifriction' properties of the 'Babbitt' and similar alloys.

A brief investigation was also made of the effects of surface finish and of continual sliding or 'running-in', and the results obtained are discussed in the light of information gained from the experiments on sliding velocity.

INTRODUCTION

The apparatus described in Part I has been used for tests upon a wide range of materials and lubricants. These experiments showed that in many cases friction is dependent upon sliding velocity, and therefore that results obtained at low sliding velocities cannot be applied directly to conditions of high sliding velocity. Accordingly, an investigation was made into the fundamental causes of change of friction with velocity, with the object of finding some logical basis for the interpretation of low velocity tests.

The friction of any combinations of materials and lubricant has also been found to be influenced by two other variables, namely, surface finish and amount of previous sliding or 'running-in'. These factors affect not only the absolute values of the coefficient of friction, but also the form of the friction/velocity relationships. It has therefore been found convenient to combine with the investigation of change of friction with velocity a limited study of these other variables.

EXPERIMENTAL METHODS

The most obvious source of changes in friction with velocity is the existence of three types of friction, dry, boundary and fluid film, for if a partial transition from one to the other occurs as a result of changing velocity, a corresponding change in friction would be expected to result. Consequently, the method of investigation adopted has been to examine the friction of different combinations of materials: first, unlubricated; secondly, with very small quantities of lubricant; and thirdly, with excess of lubricant.

The majority of tests have been made with tin-base Babbitt (containing 7 % of antimony and 3½ % of copper) running on steel and with hard steel on phosphor bronze, these systems being characteristic of the two types found in practice—soft on hard metal and hard on hard metal. Two other pairs of metals have also been made the subject of a few tests, namely, cadmium-nickel on steel and mild steel on mild steel. The sliders were prepared by dry machining, no further treatment being given. Plates were prepared by grinding with emery cloth, swabbing with caustic soda, washing with distilled water, then *iso*propyl alcohol, and drying quickly in an oven at 150° C.

Thin films of lubricant were applied to the plate by one of two methods:

(1) By deposition from a dilute solution in redistilled petroleum ether. The concentrations of solution were so chosen that 1 c.c. contained sufficient lubricant to leave a layer of approximately known molecular thickness.

Some assumptions were required for this purpose regarding the composition and orientation of the lubricants used, which were oleic acid, medicinal paraffin and I.C.E. (crankcase) oil. As regards orientation, Adam (1938) has shown that polar long-chain saturated acids when they form a condensed film on water occupy an area of about 25 sq.A per molecule, and that in this condition the molecules are closely packed with their long chains vertical.

Finch & Zahoorbux (1937) have shown that with long-chain hydrocarbons on a metal surface the molecules are also orientated vertically. It has therefore been assumed that all three lubricants orientate themselves vertically and with a spacing of 25 A per molecule. The average molecular weight of the paraffin and I.C.E. was taken as that of $C_{18}H_{38}$. This is probably considerably on the low side but would to some extent balance the error in estimating the molecular area arising from the presence of naphthenes which would be expected to occupy a greater area than corresponding paraffins.

The methods described could clearly yield only a very approximate figure for the mass of lubricant required for a given molecular thickness, but even a large error would not materially affect the conclusions to be drawn from results. For the sake of convenience, amounts of lubricant added are quoted as calculated molecular thicknesses.

The solutions described above were applied to the plate as evenly as possible from a 1 c.c. pipette, and the petroleum ether allowed to evaporate. An experiment was made to determine the effect of applying a similar quantity of the redistilled petroleum ether alone to a clean plate and allowing to evaporate. The resulting frictional value obtained was that of the clean plate.

(2) By adding excess lubricant and polishing off first with cotton-wool from which traces of grease had been removed by extraction with benzene then with ether in a Soxhlet apparatus, then with several changes of polishing cloth similarly treated.

Unless otherwise stated, experiments were made at room temperature, with a normal load of 2 kg. (total) and with a plate finished with an 0 grade emery cloth. Tests were made at five different sliding velocities, namely, 0·0096, 0·0328, 0·122, 0·526 and 2·26 cm./sec. Throughout this paper the coefficients of friction at these velocities will be referred to as μ_1, μ_2, μ_3, μ_4 or μ_5 respectively.

Experimental results

A. *Unlubricated tests*

Friction/velocity curves for unlubricated Babbitt/steel, cadmium-nickel/steel, steel/steel and hard steel/bronze are given in figure 1. Curves for fresh surfaces and also after a stated number of revolutions of the plate are shown. It will be seen that there is no significant change in friction with velocity in the case of hard steel/bronze but for Babbitt/steel and cadmium-nickel/steel (and to a lesser extent for steel/steel) there is a decrease in friction with increasing velocity. The curves for these pairs of materials become almost horizontal after running-in. Three possible explanations of the friction/velocity relationship for these latter materials may be considered.

(1) Bowden, Moore & Tabor (1943) have shown that a direct relationship exists between the observed coefficient of friction and the real area of contact between two surfaces. The real area of contact depends upon the normal load and the stress/strain characteristics of the two materials, and it is known that the stress/strain

characteristic is dependent upon the time factor. It is possible that the extent to which the one surface conforms to the changing contour of the other is greater when sliding speed is low than when it is high, so that the real area of contact, and hence the observed friction, is greater the lower the sliding speed. This effect would be expected to be much greater for the softer materials than for the harder, for the stress/strain characteristics of the former are known to be much more influenced by the time factor than are those of the latter.

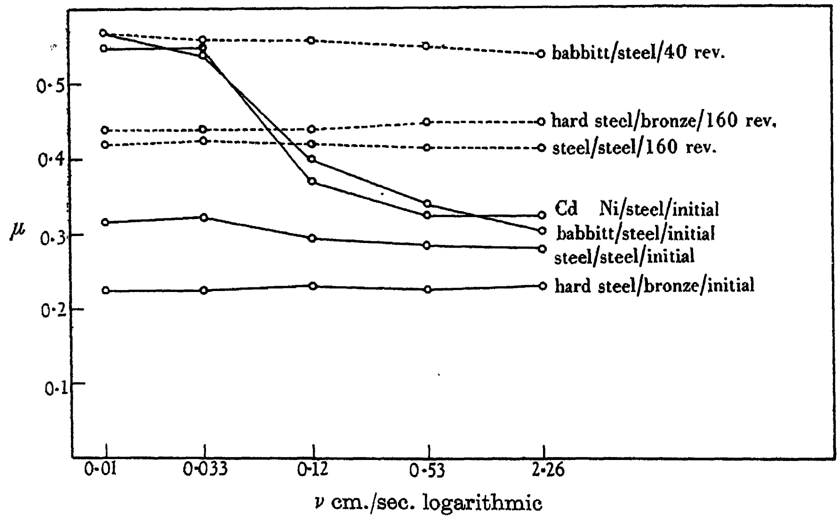


FIGURE 1. Friction/velocity curves for unlubricated systems, initially and after a stated number of revolutions.

While this explanation is at first sight attractive it leaves out of account the fact, referred to later in detail, that very small quantities of lubricant can completely eliminate the μ/ν slope. If area of contact were the controlling factor, it would be expected to be equally important whether or not the surfaces are lubricated.

(2) Unlubricated friction is commonly believed to be due to the interaction of the surface forces of the two materials. Bowden *et al.* (1943) have also shown that the junctions formed between unlubricated metals may be stronger than the metals themselves, so that shearing occurs in the metals rather than at the junction. This great strength of the junction suggests that very localized regions of solid solution or compound may have formed. For such solid solution or compound to form, diffusion over very small distances is required, in which case time will be a factor in determining the strength of the bond. In this case the greater the rate of sliding the less would be the adhesion between the two materials. While this explanation cannot be wholly discounted, it is difficult to see why the time factor should be important in the case of Babbitt/steel and cadmium-nickel/steel and not in the case of hard steel/bronze.

(3) It has been shown by Bowden & Ridler (1936) that local high temperatures may be reached during the sliding of two metals, and that in some cases the temperature may reach the melting-point of the more fusible material. The greater the rate of sliding, the greater the temperature rise to be expected, for then more work is being done at the point of contact. An increase of temperature would decrease the strength of local welds and so reduce the resistance to sliding; and if this temperature rise were highly localized, there would be no comparable increase in contact area as a result of a reduction in the yield stress of the material. The effect would be much greater for materials of low melting-point than for those of high melting-point. This seems to be the most likely explanation, although it is not immediately obvious why steel/steel should show a reduction in friction with increase of velocity, while hard steel/bronze shows no such effect. The reduction with steel/steel is, however, small, and the greater conductivity of bronze might in its case reduce the effect to insignificant dimensions.

On the basis of this hypothesis the effect of running-in in cancelling the speed effect might be derived from the greater area of contact resultant upon wear, which would decrease the pressure over the contact area, and so reduce the local temperatures.

On the whole, hypothesis (3) is the most likely of the three, but on the present evidence available it is impossible to choose with certainty between them or to exclude any other possible explanation. In this report a downward slope of the unlubricated μ/v curve will be referred to as the 'dry friction effect', no definite assumption being made as to its exact nature.

B. *Tests with lubricant films of varying thicknesses*

The usual procedure in these experiments was to place on the prepared plate a '1-molecule' or '2-molecule' film, measure μ at all five velocities, then apply increasingly thicker films up to excess of lubricant, measuring μ at each stage. By the time the last test had been made, the plate had usually made two or three complete revolutions, so results may be slightly affected by running-in effects. Such effects are normally small for such a limited amount of running-in. Plates previously used for 'dry' experiments were not, however, used without repreparation, for a single revolution, unlubricated, may exert a considerable influence on the plate.

Results may conveniently be divided into those obtained with the substantially non-polar lubricants—medicinal paraffin and I.C.E.—and those obtained with the strongly polar oleic acid.

Paraffin and I.C.E. Results for Babbitt/steel are shown in figure 2 (with paraffin) and figure 3 (with I.C.E.). The addition of a 2-molecule film of paraffin greatly reduces μ below that for the unlubricated surfaces (see figure 1) and almost eliminates the downward slope. Increasing the thickness of the paraffin film up to 8 molecules further reduces the coefficient. Between 8 and 20 molecules the curve obtained is constant and has a slight upward slope. With 200 molecules there is a slight reduction in μ_1, and a considerable reduction in μ_5. With 2000 molecules and excess there is no change in μ_1 but a further reduction in μ_5. Thus with increasing thicknesses of lubricant film, the μ/ν curve changes in two stages: first a reduction in μ at all velocities and a flattening of the curve till it becomes horizontal; secondly a reduction in μ_5, with little or no change in μ_1, so that the curve again develops a downward slope.

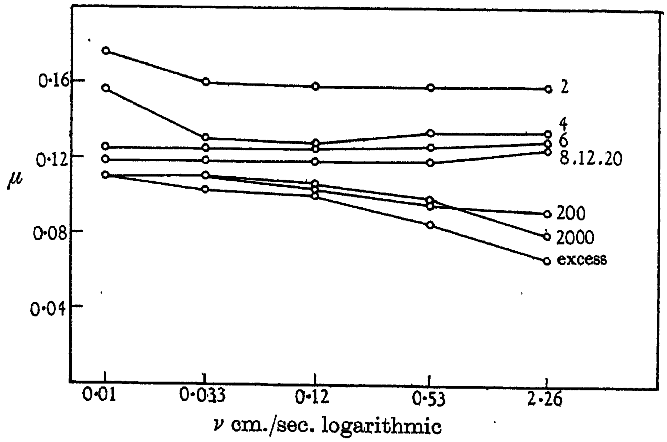


FIGURE 2. Babbitt/steel with solvent-deposited films of paraffin. In this and all subsequent diagrams, numbers by curves refer to approximate molecular thickness of lubricant film.

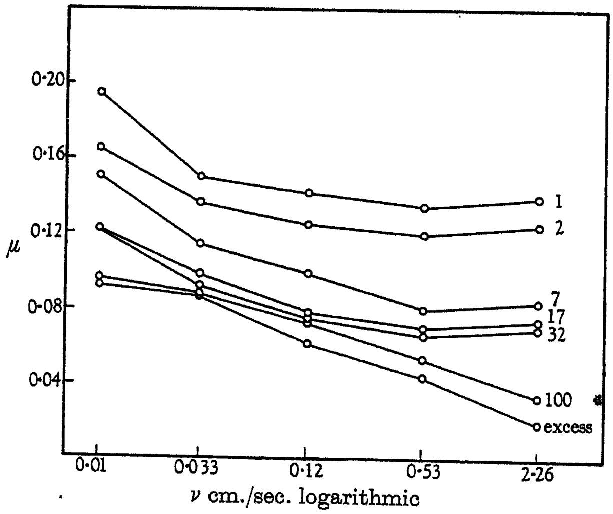


FIGURE 3. Babbit/steel with solvent-deposited films of I.C.E.

With I.C.E. the results observed are essentially similar, but the two stages appear to have merged, so that the curve never becomes entirely horizontal. It is interesting to note that under these conditions all the thin-film curves show a minimum at speed 4 (0·52 cm./sec.).

Results for hard steel/bronze with paraffin are shown in figure 4. In this case with the thinnest films the μ/ν curve has a marked upward slope, but with increasing amounts of lubricant it tends to become horizontal. No experiments were made with solvent-deposited I.C.E.

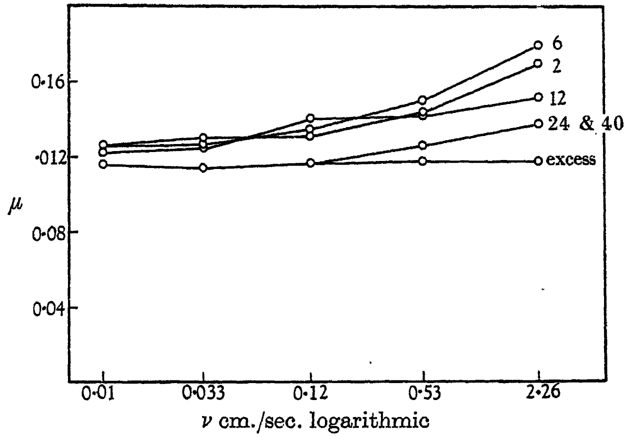


FIGURE 4. Hard steel/bronze with solvent-deposited films of paraffin.

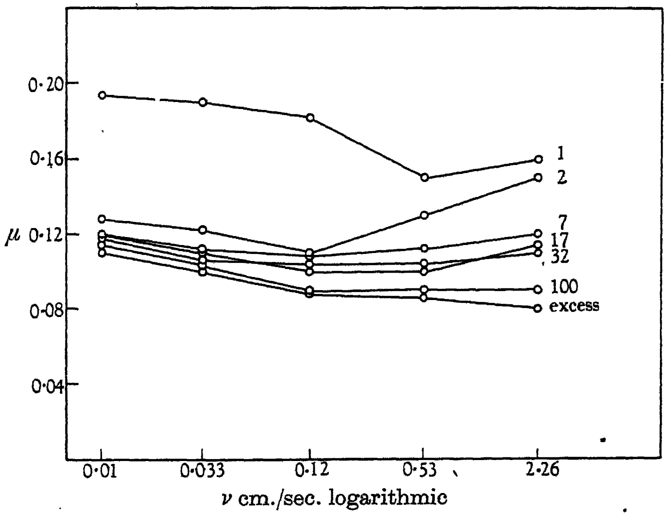


FIGURE 5. Cadmium-nickel/steel with solvent-deposited films of I.C.E.

The curves for cadmium-nickel/steel with I.C.E. shown in figure 5 are interesting, in that with thin lubricant films there is a downward slope at the low velocities and an upward slope at the higher velocities, with a minimum at an intermediate velocity.

With steel/steel and I.C.E. (figure 6) entirely different results were obtained. The μ/ν curve which is substantially horizontal when the specimens are unlubricated, develops a marked downward slope with 5- and 10-molecule films, becoming substantially horizontal again with increasing film thickness up to excess. With the much rougher surface obtained by machining, a downward-sloping curve is retained even with excess. The μ/v curve with a machined surface and excess of I.C.E. is shown as an interrupted line in figure 6.

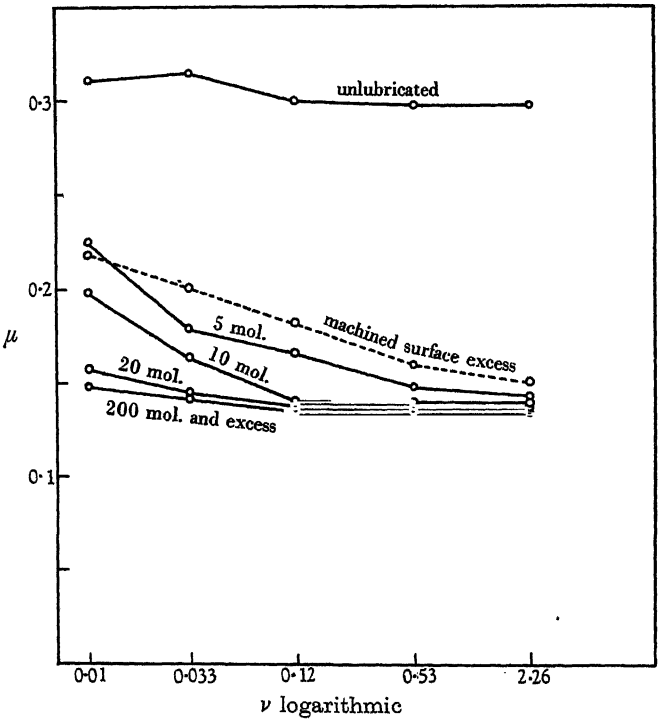


FIGURE 6. Steel/steel with solvent-deposited films of I.C.E.

Oleic acid. With this lubricant similar effects are observed for Babbitt/steel, steel/steel and for hard steel/bronze (figures 7, 8 and 9 respectively). With a 2-molecule film the μ/v curves have a marked upward slope. With a 6-molecule film there is a slight reduction in μ at all velocities, with no significant change in the slope of the curve. With more lubricant up to excess there is no further change, except that μ at high speed for Babbitt/steel drops slightly.

Discussion

(1) Before proceeding to a general hypothesis interpreting these results, it may be profitable to deal with one observation, which appears curious at first sight. With paraffin and I.C.E., and to a much smaller extent with oleic acid, there are significant differences between values of μ observed with different nominal thicknesses of film, over ranges of thickness in which fluid film lubrication is at least very unlikely. It is considered certain that these differences do not imply that boundary films between the surfaces can change in thickness according to the

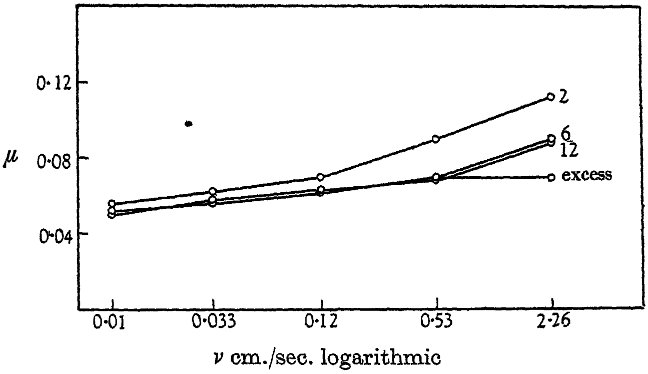


FIGURE 7. Babbitt-steel with solvent-deposited films of oleic acid.

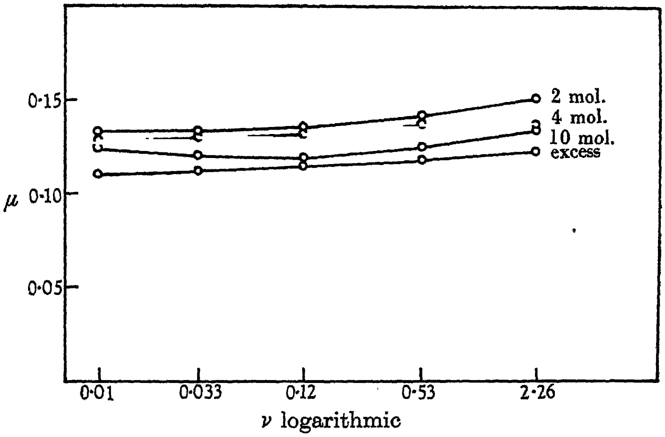


FIGURE 8. Steel/steel with solvent-deposited films of oleic acid.

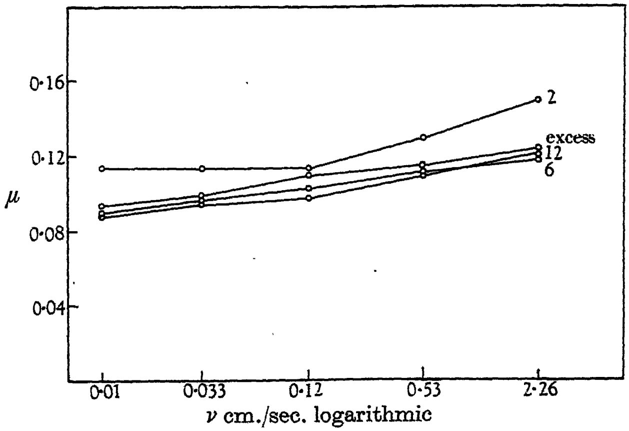


FIGURE 9. Hard steel/bronze with solvent-deposited films of oleic acid.

quantity of lubricant present, but that the differences are derived from the method adopted to apply the lubricant. If the lubricant solution were perfectly evenly spread and the surface perfectly flat, then an even film of lubricant should result on the evaporation of the solvent. But the prepared surface has, in fact, hills and valleys of large magnitude in terms of molecular dimensions. When the greater part of the solvent has evaporated, the remainder will tend to collect in the valleys, leaving the hills less well covered. When all the solvent has evaporated the majority of the lubricant molecules will be left in the valleys. This segregation will be opposed by the tendency of the lubricant molecules to attach themselves to the plate surface even while sufficient solvent is present to hold them in solution, which tendency will be large for strongly polar lubricants and small for non-polar ones. Thus when a 2-molecule film is referred to it should be considered to imply that a mass of lubricant sufficient to provide such a film is present, and not that such a film is actually in being. This explanation is strongly supported by tests with wiped films and with different surface finishes to be dealt with later. The solvent method of applying the lubricant can thus be considered to have achieved its object of providing transition conditions between dry and boundary friction.

(2) The results obtained with Babbitt/steel and paraffin (figure 2) showed that changes in the μ/ν curve with an increasing amount of lubricant take place in two stages. The first stage in which the μ/ν curve becomes increasingly horizontal, and the dry frictional properties increasingly masked, is satisfactorily explained by the foregoing hypothesis, and on this basis it appears that a complete boundary film is established when a nominal 8-molecule layer has been deposited on the plate. The characteristic of the second stage is that there is little change in μ_1 with increasing amount of lubricant, but a considerable change in μ_5, up to and beyond a 2000-molecule layer. The same change in μ_5 with increasing film thickness is found with Babbitt/steel and I.C.E. (figure 3) and to a considerably less extent with oleic acid. There is also some evidence of such a change with hard steel/bronze and paraffin (figure 4). There can be little doubt that this second stage effect is due to the onset of fluid film lubrication. Small irregularities in the surfaces provide the opportunity for minute oil wedges to form and support part of the load; the higher the speed the greater is the magnitude of this effect.

This explanation was confirmed by arranging an electrical circuit so that a p.d. of 40 μV was placed between one slider and the plate. A galvanometer in the circuit indicated the current passing. This showed a steady constant deflexion at low velocities, but with Babbitt/steel and I.C.E. it showed fluctuations at 0·52 cm.sec. which became more marked as the velocity increased.

(3) It has been shown that there are two possible sources of a slope in the μ/ν curve—the properties of the unlubricated system and the transition from boundary to fluid film lubrication. Both, however, result in a decrease in μ with increasing ν, whereas a number of μ/ν curves have been obtained showing an increase in μ with increasing ν. There must, therefore, be some effect which results in a change in the latter direction. It is hardly likely that a transition from fluid to boundary

lubrication can occur with increase of velocity, but a transition from boundary to dry friction is much more probable. This latter possibility is suggested by the fact that for hard steel/bronze and cadmium-nickel/steel thin films of I.C.E. give upward sloping curves, the slope of which is diminished by the addition of more lubricant. This strongly implies a sparse-lubrication effect. It has been shown by Bowden & Leben (1940) by Dacus, Coleman & Roess (1944) and by others that a lubricant film is partially destroyed during sliding. Since the specimens are in contact over a finite area, when sliding is taking place under boundary conditions the rear end of each slider must be in contact with a part of the plate from which the lubricant film has partially been removed by contact with the front end. It is improbable therefore that a complete film is present during sliding on the whole of the contact area. A higher value of μ than the true one will thus normally be observed. The extent of this increase will depend upon the rate of breakdown of the lubricant film. It is reasonable to suppose that volatilization and thermal decomposition play a large part in this breakdown, which will therefore be greater as the sliding speed increases. Thus with increasing sliding speed an increasing tendency may be expected for boundary friction to be replaced by dry friction over part of the contact area, and a higher value of μ accordingly observed. In this paper the effect will be referred to as the 'film destruction' effect.

(4) To summarize the above hypothesis, it is suggested that an increase in velocity of sliding may result in:

(i) A decrease in μ due to some effect dependent upon physical properties of the unlubricated system.

(ii) An increase in μ due to a transition from boundary to dry friction.

(iii) A decrease in μ due to a transition from boundary to fluid friction.

It is quite conceivable that all three effects may occur simultaneously, for the local pressure may vary widely over the contact area. At one point it may be sufficiently high to break down the lubricant film, while at another it is low enough to allow a fluid film to form. The observed μ/v curve will be the resultant of the three effects.

(5) Assuming this hypothesis to be substantially correct, consider the effect of adding increasing quantities of lubricant to the three systems Babbitt/steel, steel/steel and hard steel/bronze, as expressed in figures 2–9.

Babbitt/steel. With thin films, two tendencies act in opposition and the observed shape of the curve is the resultant of the two. With paraffin the 'dry-friction' effect is predominant with 2- and 4-molecule films, but with 6–20-molecule films the 'dry-friction' effect and the 'film-destruction' effect almost exactly cancel each other. With I.C.E. the 'dry-friction' effect predominates, except at the highest velocity, so that a minimum in the curve results. With oleic acid (figure 6) probably owing to the formation of a more complete boundary film, the 'film-destruction' effect predominates, and curves with an upward slope result.

With greater quantities of lubricant, fluid film effects occur with all three lubricants, but especially with the more viscous I.C.E. and paraffin. With paraffin there is little evidence of fluid film effects below 0·12 cm./sec. (speed 3). With I.C.E. such effects certainly occur above this velocity, but it is difficult to decide whether the μ/v slope below this velocity is due to fluid film formation or to the remnants of a 'dry-friction' effect. With oleic acid the only evidence of fluid film formation is in the difference at the highest velocity between a 6-molecule film and excess.

Hard steel/bronze. The μ/v characteristics for paraffin and oleic acid are very similar. With both lubricants, very thin films result in a μ/v curve with a marked upward slope (there being no 'dry-friction' effect to neutralize it). This slope is reduced by increasing the film thickness, possibly because the more complete the boundary film the lower is the temperature at the interface and hence the lower the rate of film destruction. There is little evidence of fluid film formation with this pair of materials, except possibly for μ_5 with paraffin, when there is a difference between a 40-molecule film and excess.

Steel/steel. The behaviour of this system with I.C.E. differs from that of the others in that the downward slope of the μ/v curve, slight for the unlubricated surfaces, is increased by certain conditions of lubrication. These conditions represent imperfect boundary lubrication, either by reason of the very small amount of lubricant present or of a very rough finish. (As shown later in this paper, a rough surface predisposes to imperfect boundary lubrication.) It may be that the effect of such conditions of lubrication is to enhance the dry friction effect in some way, but it is also possible that with this system there is a fourth source of change of coefficient with velocity, and further investigation of this subject is clearly required.

C. *Experiments with rubbed films*

A few experiments were made with rubbed films, with the following objects:

(*a*) To examine the hypothesis that incomplete films may result from deposition from solvents, even though sufficient lubricant for a complete film is present.

(*b*) To obtain confirmatory evidence regarding fluid film formation.

(*c*) To examine whether a rubbed film is identical with one produced by forcing out lubricant under pressure, as occurs in practice.

In the first place, a 2-molecule layer of I.C.E. was deposited on a steel plate, and μ against Babbitt measured. The plate was then rubbed with a degreased cloth and again tested. The results (figure 10) show that there was a significant fall in μ, implying a better distribution of the lubricant. This tends to confirm the hypothesis previously stated regarding deposited films.

Secondly, excess of I.C.E. was dropped on to a steel plate and the friction against Babbitt measured. The excess was then removed with degreased cotton wool and polished with several changes of degreased polishing cloth. The friction was again measured (figure 10). There was a rise in μ_3, μ_4 and μ_5, but no change in μ_1 and μ_2. Excess was again added, and the original curve obtained. This indicates that with this lubricant there is a tendency to fluid film formation at the higher velocities. This experiment was then repeated with oleic acid, but in this case no change was observed in μ at any velocity on substituting a rubbed film for excess. This result may be attributed to two causes. First, the boundary coefficient is lower with oleic acid than with I.C.E., so less difference is to be expected between the friction with a boundary film and that with a thin fluid film; secondly, oleic acid has a lower viscosity than I.C.E., so fluid films will form less readily.

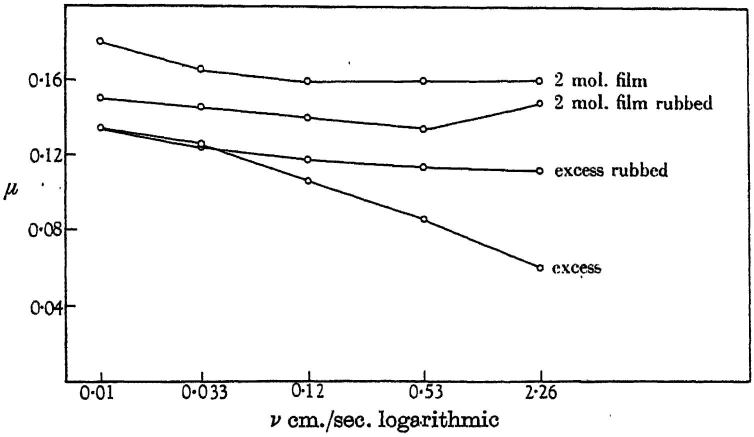


FIGURE 10. Babbitt/steel with solvent-deposited and rubbed films of I.C.E.

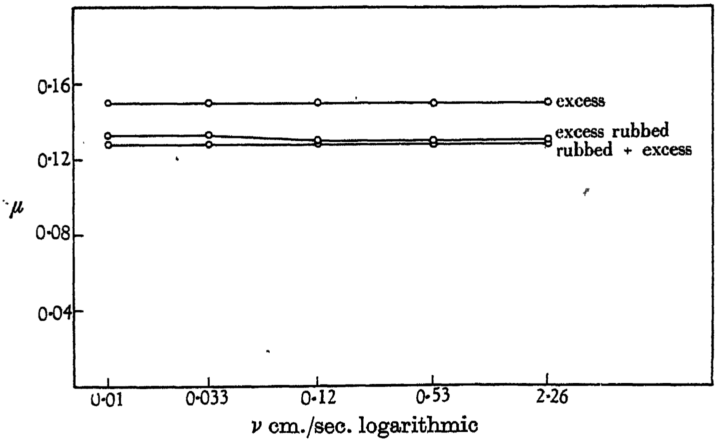


FIGURE 11. Hard steel/bronze with rubbed films of I.C.E.

A similar experiment was made with hard steel/bronze and I.C.E. The results are expressed in figure 11. In this case rubbing brings about a significant decrease in μ; the addition of excess to the rubbed film brings about no further change. This experiment demonstrates two things. First, that rubbing can bring about an improvement in the frictional properties of a lubricant film, probably as a result of an improvement in its orientation. Secondly, that the combination hard steel/bronze/I.C.E. shows (with the specimen geometry used in this apparatus) no tendency to fluid film formation at speeds up to 2·26 cm./sec., whereas with Babbitt/steel/I.C.E. such a tendency is present at speeds above 0·12 cm./sec.

The confirmation given by rubbed-film experiments of the very marked tendency of the combination Babbitt/steel to fluid film formation is of considerable interest. It appears highly probable that this characteristic is responsible for the known good 'anti-friction' properties of white metals as compared with harder alloys, for, as experiments reported here show, the true boundary coefficient of Babbitt/steel is little or no lower than that of hard steel/bronze. A series of experiments was made to determine whether this property is associated with the yield stress of the metal. Sets of sliders were made of pure tin and 1, 2 and 3 % antimony in tin respectively. The alloys were annealed to ensure that those containing antimony consisted of homogeneous solid solutions. These sliders were then tested against a steel plate, firstly with excess of I.C.E. and secondly with a rubbed film of I.C.E. The μ/v curves obtained with the different sets of sliders are shown in figure 12. It will be seen that as the hardness of the sliders increases, the difference between the μ/v curves for excess and for a rubbed film (representing the tendency to fluid film formation) diminishes. In figure 13 the difference between the values of μ_5 for excess and for rubbed films is plotted against Vickers Hardness, and a rough correlation is observable. This relationship is not maintained for the Babbitt sliders (see figure 10), which show a difference in μ_5 values of 0·052, although their hardness is about 25. This must be attributed to the two-phase structure of this alloy (solid solution + Cu_6 Sn_5).

The dependence of the fluid film forming ability of a material upon its hardness is probably due to the relationship between real area of contact and yield stress. The area of actual contact which is established between two stationary specimens is also the area over which minute oil wedges can form when the specimens are in relative motion. The greater this area, the greater the number and size of such oil wedges, and hence the greater the proportion of the load they can carry. Thus the lower the yield stress of the softer material of a pair the greater the proportion of the load carried by oil wedges.

The oil pressures in such wedges depend not only upon the speed of sliding but also upon the thickness of the oil film. If one surface is very rough, then high spots will come in contact with the other surface while other parts of the surfaces are separated by an oil film too thick to develop a considerable pressure. Slight surface irregularities are, of course, essential to provide the necessary 'wedges', but roughness in excess of this requirement will reduce the proportion of the load carried by oil films. Experiments, to be described later in this paper using different surface finishes, show that this factor has, in fact, a very marked effect upon the coefficient of friction when there is a tendency to fluid film formation.

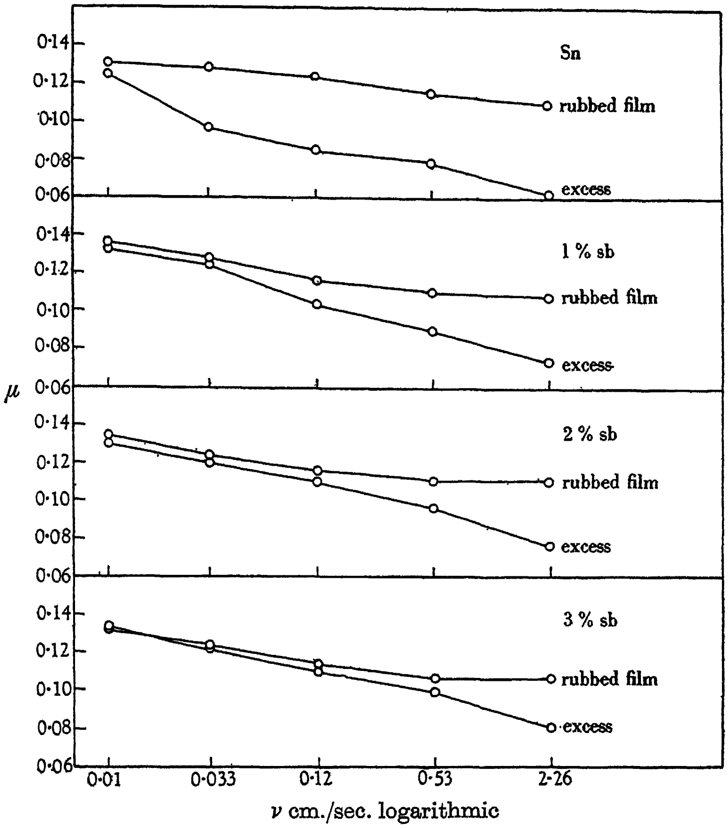


FIGURE 12. Friction/velocity curves with rubbed films and with excess of lubricant for tin and tin alloys containing 1, 2 and 3 % of antimony.

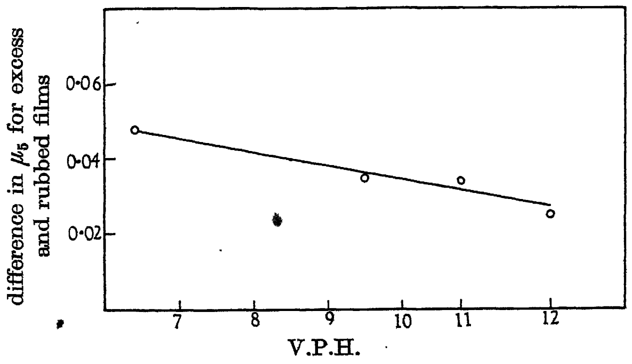


FIGURE 13. The influence of hardness of homogeneous tin-antimony alloys on the tendency to fluid film formation. The difference between μ_5 with excess of lubricant and μ_5 with a rubbed film is plotted against V.P.H.

D. *The effect of other factors*

(1) *Running-in with lubricated surfaces*

The results so far given for lubricated surfaces all refer to tests made on fresh surfaces, or with not more than one or two revolutions of the plate, each revolution corresponding to a repetition of sliding over the same track. With a lubricant present the effect of one or two revolutions is small, but progressive, and after a considerable number of revolutions becomes very significant. The practice was adopted in each experiment of making all required 'fresh-surface' measurements, then allowing the plate to make 160 revolutions at 2·26 cm./sec. (the maximum velocity used) with the customary normal load of 2 kg. and excess of lubricant. The procedure is here referred to as 'running-in' though the use of this phrase is not intended to imply that the changes observed are necessarily analogous to those which occur in running-in a bearing.

In all the cases examined, such running-in has been observed to increase the downward slope of the μ/v curve. If the curve has initially a downward slope, the effect is to make this slope steeper; if the curve has initially an upward slope, it is made horizontal or given a downward tendency. Usually μ_1 is increased and μ_5 decreased, but this is not invariably the case, for sometimes the general level of the curve is changed.

Some tests were made to find what degree of repreparation was necessary to remove the effects of running-in. In figure 14, μ/v curves for Babbitt/steel with excess of paraffin are shown. Curve *A* was obtained initially, curve *B* after 160 revolutions. It will be seen that running-in has greatly increased μ_1 while μ_5 is unchanged. The plate and sliders were then rinsed thoroughly with petroleum ether and dried. A 20-molecule film was then placed on the plate and a further test made (curve *C*). μ_1 has dropped, but μ_5 has increased, since fluid film formation is now restricted. Excess of lubricant was then placed on the plate and curve *D* obtained. This curve shows that a substantial change in the μ/v curve can be obtained simply by dissolving off the lubricant and replacing with fresh. Lastly, the plate was reprepared in the usual way, excess paraffin added and curve *E* obtained, substantially the same as with fresh sliders. A second and similar experiment showed that the original 'fresh-surface' curve could be partly restored by rubbing the plate hard with a polishing cloth.

Similar effects were observed with hard steel/bronze and I.C.E. In figure 15 curve *A* is the initial curve, and curve *B* that after 160 revolutions. The plate was then rubbed hard (circumferentially) with polishing cloth, and curve *C* (similar to curve *A*) obtained. If instead of rubbing the plate, the old lubricant is dissolved off and replaced by fresh, a curve intermediate between *A* and *B* results.

These tests demonstrated that the 'running-in' effect is partly at least attributable to changes in the lubricant film, since initial values of μ can be partially or completely restored by operations which cannot substantially affect the metal surfaces. In the case of hard steel/bronze, rubbing with a cloth can almost completely restore initial values of μ (figure 15, *C*), but with Babbitt/steel, the curve obtained after such rubbing is still significantly different from that obtained initially. It is probable that in the latter case some change in the metal surfaces is also involved, possibly transfer of Babbitt to the steel plate.

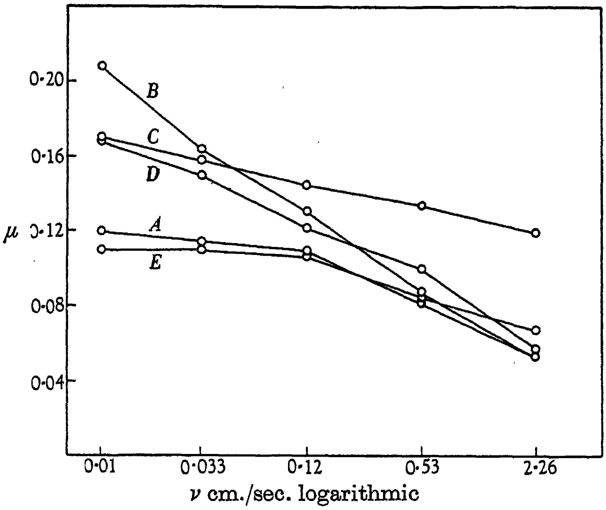


FIGURE 14. Babbitt/steel with paraffin. *A*, initial with excess of paraffin. *B*, after 160 revolutions excess of paraffin. *C*, plate and sliders cleaned with petroleum ether, 20-molecule film added. *D*, as *C* + excess of paraffin. *E*, plate reprepared.

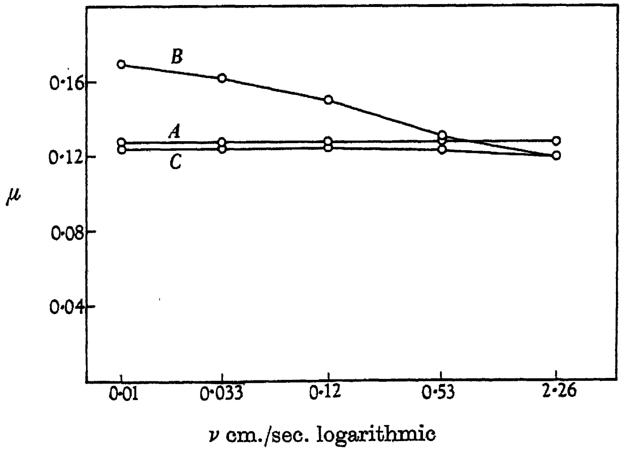


FIGURE 15. Hard steel/bronze with I.C.E. *A*, initial with excess of I.C.E. *B*, after 160 revolutions excess of I.C.E. *C*, as *B*, plate rubbed with polishing cloth.

The reduction in μ_5 which frequently occurs on running-in can be satisfactorily explained by the flattening of the sliders by wear, which provides greater opportunity for fluid film formation. The tendency for μ_1 to be increased cannot be thus explained. Two experiments were therefore made to determine whether this increase was connected with the sliding velocity during running-in. The systems used were hard steel/bronze/I.C.E. and Babbitt/steel/I.C.E., and both were tested before and after running-in at 0·01 cm./sec. Results are shown in figure 16. With Babbitt/steel (curves X and Y) the effect of running-in is much the same as at the normal velocity (2·26 cm./sec.), but with hard steel/bronze (curves A and B) the effect is entirely different. In place of the increase in μ_1 and decrease in μ_5, there is a decrease in μ at all velocities. It appears therefore that in the case of hard steel/bronze, the effect of running-in at μ_5 is so to influence the lubricant film, possibly by changing its molecular orientation, that its low-speed frictional properties are adversely affected. It may be noted that this is the only case that has been observed of a change in coefficient with velocity not attributable to transition from one type of friction to another.

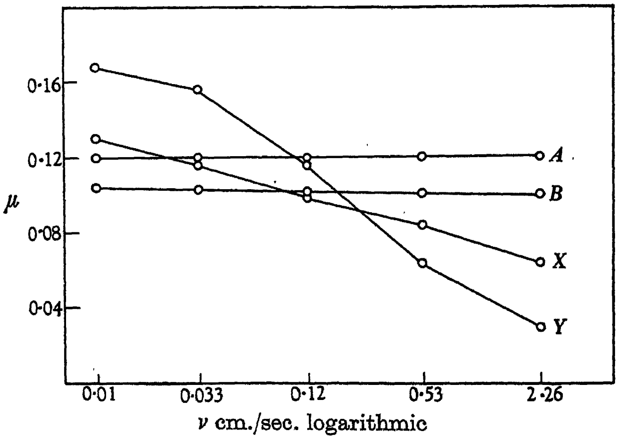


FIGURE 16. Babbitt/steel and hard steel/bronze with I.C.E.

A, initial with excess of I.C.E. } Hard steel/bronze.
B, after 320 revolutions at ν_1 (0·01 cm./sec.) }
X, initial with excess of I.C.E. } Babbitt/steel.
Y, after 320 revolutions at ν_1 (0·01 cm./sec.) }

(2) *Surface finish*

The standard surface finish of the plate was that given by an 0 grade emery cloth. A few experiments were also made with plates finished by means of a 1 grade emery cloth, and one isolated test was made with a plate polished on a buffing wheel to a mirror finish.

With unlubricated surfaces no significant difference was found between curves obtained with a 1 emery finish and those with an 0 emery finish, either for Babbitt/steel or for hard steel/bronze. The two curves for Babbitt/steel are shown in figure 17.

With Babbitt/steel lubricated with paraffin, the surface finish was found to exercise a considerable influence. Figures 18 and 19 show the curves obtained for this system with different thicknesses of lubricant film, figure 18 with a polished plate and figure 19 with a 1 emery finished plate. (Results for the standard finish were given in figure 2). With a polished plate, the two stages of change in the μ/ν curve are clearly separated. Stage 1 is complete when a 6-molecule film is present, and stage 2 commences with a 12-molecule film, a thinner film than that required with the standard finish. This is in accord with the hypothesis that stage 2 is the result of a transition from boundary to fluid film lubrication, for a smooth surface may be expected to favour the formation of fluid films. With the 1-emery finished plate the two stages merge, no horizontal curve being obtained. Values of μ are very much higher than with the polished or 0 emery finished plate. Running-in tends to emphasize the differences between the three finishes. Figure 20 shows μ/ν curves for the three finishes with excess of paraffin, initially and after 160 revolutions. (In this figure, and figures 19 and 20, initial results are shown by interrupted lines.)

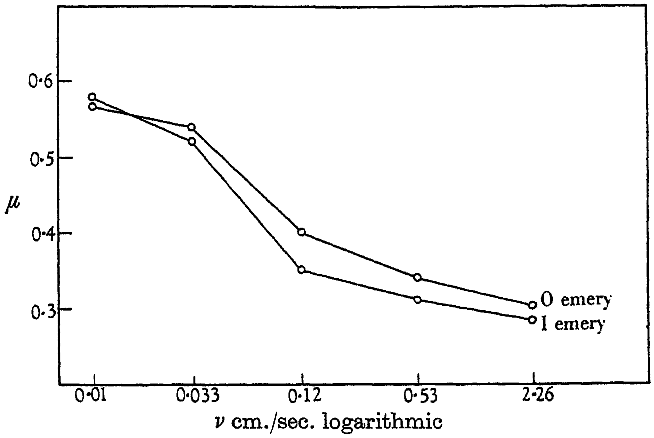


FIGURE 17. Babbitt/steel, unlubricated, with two different finishes of plate.

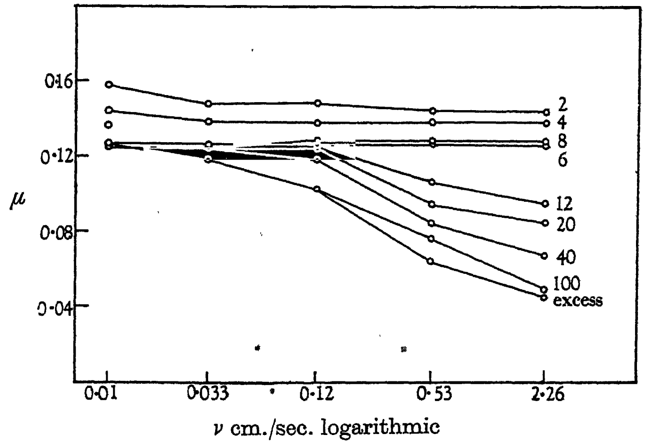


FIGURE 18. Babbitt/steel with polished steel plate and solvent-deposited films of paraffin.

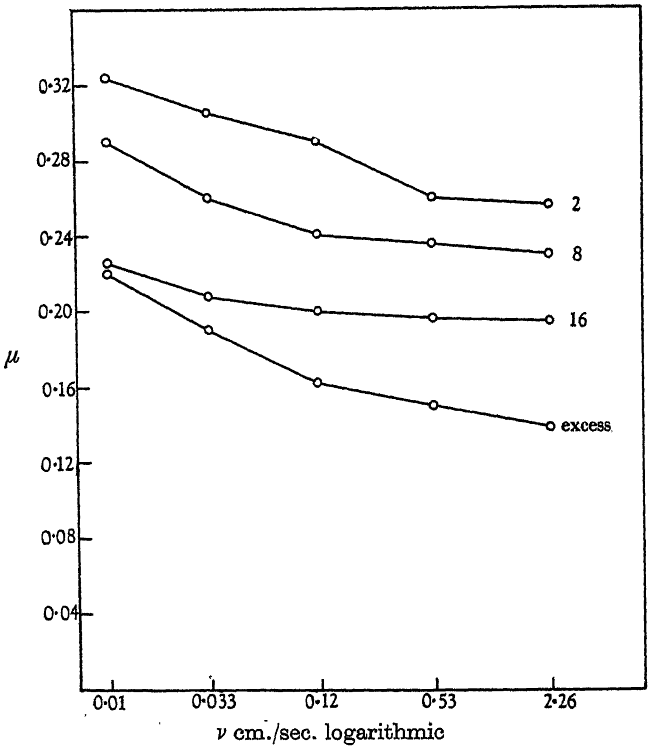


FIGURE 19. Babbitt/steel with coarse-emery finished steel plate and solvent-deposited films of paraffin.

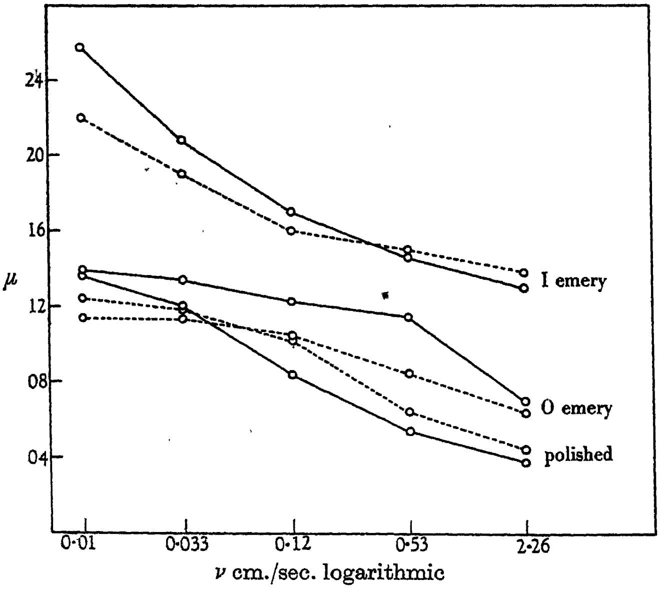


FIGURE 20. Babbitt/steel with excess of paraffin, showing initial curves (interrupted lines) and curves after 160 revolutions (full lines) with three different finishes of plate.

With oleic acid as lubricant, surface finish has less effect on the μ/v curves for Babbitt/steel. Figure 21 shows μ/v curves for this system initially and after 160 revolutions for standard and 1-emery finished plates with excess of oleic acid. There is no significant difference in the initial results, but running-in causes a greater rise in μ in the case of the 1-emery finish.

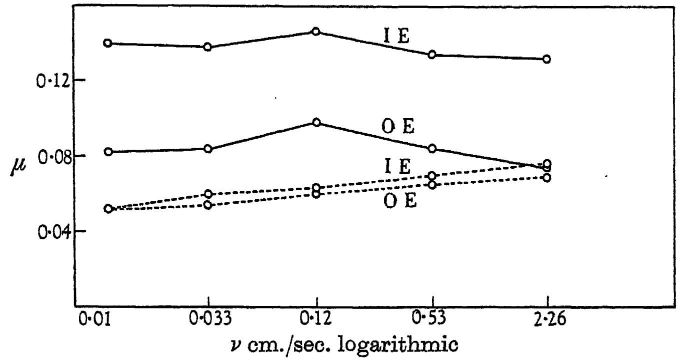


FIGURE 21. Babbitt/steel with excess of oleic acid, showing initial curves (interrupted lines) and curves after 160 revolutions (full lines) with two different surface finishes of steel.

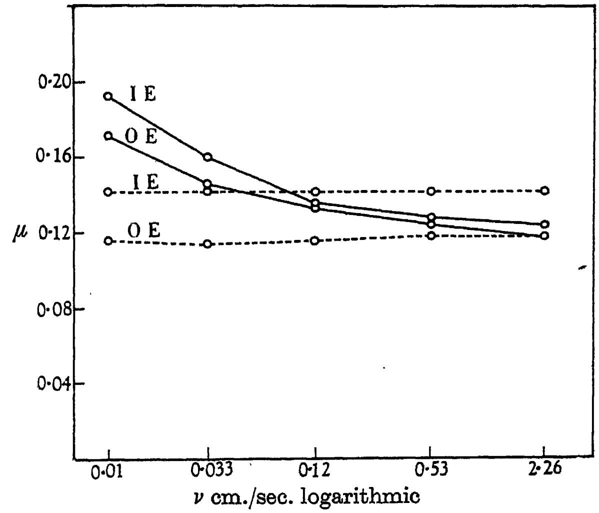


FIGURE 22. Hard steel/bronze with excess of paraffin, showing initial curves (interrupted lines) and curves after 160 revolutions (full lines) with two different surface finishes of bronze.

With hard steel/bronze, surface finish of the plate has less influence than with Babbitt/steel. With both paraffin and oleic acid as lubricants the rougher surface yields rather higher values of μ both initially and after 160 revolutions. The curves for two different finishes, with paraffin as lubricant, are shown in figure 22.

Since surface finish has no influence on the friction of the unlubricated surfaces, its influence on lubricated surfaces must be due to an effect on the nature of the lubricant film. In the first place a rough surface may be expected to delay the onset of fluid film lubrication. This effect has been observed by Beeck, Givens & Smith (1940), who showed, by means of the four-ball top apparatus, that a polished surface is much more conducive to fluid film formation than a rough one. In addition Lewis (1940) has shown that the minimum point of the $\mu : \frac{zn}{p}$ curve for a journal bearing is reduced by improved surface finish. Thus when the conditions of speed, pressure and lubricant viscosity are such that some degree of fluid film formation is possible (as with Babbitt/steel, I.C.E. and higher velocities), lower values of μ should be observed with smoother surfaces. Examination of the results recorded shows that this is, in fact, the case. The results also show, however, that the surface finish also exerts an influence under conditions when fluid films are substantially absent (e.g. at the lowest velocity). A possible reason for this is suggested by the observation of Finch & Zahoorbux (1937) that the perfection of orientation of the hydrocarbon film on copper is greatly influenced by the surface finish of the copper. This explanation is supported by the fact that μ is much more influenced by surface finish with the essentially non-polar lubricants paraffin and I.C.E. than with the strongly polar oleic acid. The non-polar molecules must be orientated by reason of lateral forces, whereas polar molecules are orientated by the attraction for the metal surface of the polar end-group. The latter form of orientation is likely to be much more capable of resisting the disturbing action of an uneven surface than the former.

If this explanation is the true one, the relative insensitivity of hard steel/bronze to the surface finish of the latter material may possibly be explained by the fact that its surface contour at the region of contact will, in any case, largely be controlled by the finish of the steel, which is the harder material.

SUMMARY OF RESULTS AND CONCLUSIONS

The observed coefficient of friction for a given combination of materials and lubricant depends upon:

(1) *Sliding velocity*

The systems examined have all shown marked changes in friction with velocity between the limits 0·01–2·25 cm./sec. These changes may be derived from at least three sources.

(*a*) From the properties of the clean metal surfaces. Certain combinations of materials show a decrease in friction with increasing velocity when unlubricated. The physical property or properties controlling this tendency have not yet been definitely identified but the fact that the tendency is most marked when one component is a soft metal of low melting-point suggests that it may be associated with reduction in shear strength of the metal with increasing temperature.

(*b*) From the transition from boundary to fluid film conditions. This tendency is also most marked when one component is a soft metal, in which case the real area of contact is greater and the local pressures lower than for two hard metals. It results, of course, in a decrease of friction with increasing velocity.

(*c*) From some factor present under boundary conditions which makes for an increase in friction with increasing velocity. There is evidence to suggest that this increase is due to a partial destruction of the boundary film, the rate of which destruction rises with increasing sliding speed.

(2) *Quantity and method of application of lubricant*

When insufficient lubricant is present for a complete boundary film to be formed, frictional characteristics intermediate between dry and boundary conditions are found. On the other hand with excess of lubricant, fluid films may tend to form. There is evidence that the frictional properties of lubricant films produced by rubbing differ from those of films produced by squeezing out excess under pressure.

(3) *'Running-in'*

Running-in brings about changes in the physical condition and geometry of the metal surfaces and also in the state of the lubricant film. In some cases, a change in the metal surfaces, possibly transfer of material, causes an increase in friction, but this may be offset as regards the higher velocities by more favourable conditions for fluid film formation resulting from a flattening of the sliders by wear. Running-in at a high velocity has been shown in one instance so to influence the lubricant film as to increase the low-velocity friction.

(4) *Surface finish*

For the combinations of metals tested, the influence of surface finish is wholly through its effect on the lubricant film, for the friction of the unlubricated surfaces is not dependent on surface finish. First, a rough surface delays the onset of fluid film conditions. Secondly, a substantially non-polar lubricant forms on a rough surface a boundary film less effective than on a smooth surface. It is suggested that this may be because the rough-surface film is less perfectly orientated. Thirdly, the effect of running-in on the lubricant film is much greater when the plate surface is rough than when it is smooth.

The surface finish of the plate has much more influence in the combination Babbitt sliders/steel plate than in the combination hard steel sliders/bronze plate. This is not unexpected, for in the latter case the contour of the contact area will be controlled by that of the steel surface rather than by that of the bronze.

PRACTICAL IMPLICATIONS

In considering circumstances in which boundary lubrication may occur in practice, two distinct cases are to be considered. First, the fluid film may be absent because the balance between supply and loss of lubricant is such that insufficient lubricant is present to maintain such a film. Such circumstances may occur, for example, in a badly designed bearing, in which oil escapes from regions of high pressure by oil grooves, or in any sliding mechanism which has no positive mechanical oil feed. The factors which are then important are the boundary coefficient of friction and the resistance of the boundary film to breakdown during sliding. The latter property has been tested directly by Dacus, Coleman & Roess (1944), and could also be tested directly with the Tin Research Institute apparatus. On the present hypothesis some indication of this property may be obtained from the tendency of the μ/v curve to show an upward slope. For example, oleic acid has a much greater tendency than have paraffin or I.C.E. to give μ/v curves with an upward slope, and may therefore be no better as a lubricant under conditions of sparse supply, even though its boundary coefficient is lower.

Secondly, boundary lubrication may occur as a result of high local pressures, low sliding speeds and low lubricant viscosity, under conditions of ample supply of lubricant. The conditions giving rise to boundary lubrication with ample lubricant present have been shown by the experiments described to vary widely with different materials. With Babbitt sliding on steel at 2·26 cm./sec. fluid film formation reduces the observed coefficient to about one-third of its 'boundary' value, whereas with hard steel on bronze no fluid film formation is observable at this velocity. The capacity of a pair of materials to form a fluid film may well be a much more important property than their boundary coefficient in many practical circumstances. A high capacity for fluid film formation narrows the range of operating conditions under which boundary friction and consequent damage to the surfaces can occur. The present work has been mainly concerned with two very different pairs of materials, but the same or a similar technique might well yield valuable comparisons between different bearing materials, and also serve to indicate the degree of importance to be attached to other variables, e.g. surface finish.

Under conditions of ample lubrication, it is doubtful if the resistance of the boundary film to breakdown is a very important factor, for areas of real contact between sliding surfaces are usually very small, so ample opportunity is afforded for film repair.

In both circumstances, sparse or ample lubrication, the effect of running-in on the properties of a combination of materials and lubricant are important as regards both effect upon lubricant film or upon the surfaces.

It appears, therefore, that the frictional properties of a combination of materials can be defined by the following:

(1) The unlubricated μ/v curve for the two materials.

(2) The resistance to breakdown of the boundary film. ((1) and (2) are only of importance with sparse lubrication.)

(3) The low-speed value of μ, representing an approximation to the true boundary coefficient of friction.

(4) The effect of running-in on the low-speed value of μ, and whether this is due to change in lubricant film or to an effect upon the materials. In the former

case the increase is only important when high-speed sliding is succeeded by low-speed. In the latter case, a change corresponds to a change in the true boundary coefficient.

(5) The tendency to fluid film formation (this might be defined by the sliding velocity at which fluid film effects become apparent with a given specimen geometry) and the effect of running-in upon this tendency.

(6) The effect of surface finish upon the low-speed value of μ and upon the transition from boundary to fluid film lubrication. The relative importance of these effects will depend upon the practical circumstances.

This work was carried out at the Tin Research Institute and formed part of a research programme which was undertaken for the Joint Admiralty and Ministry of Supply Advisory Service on Lubrication and financed by the Ministry of Supply. The authors are indebted to the Director General of Scientific Research (Defence) for permission to publish the results, and also wish to express their gratitude to Dr D. Clayton for helpful advice at several stages of the research.

REFERENCES

Adam, N. K. 1938 *The physics and chemistry of surfaces*, p. 30. Oxford: Clarendon Press.

Beeck, O., Givens, J. W. & Smith, A. E. 1940 *Proc. Roy. Soc.* A, **177**, 91.

Bowden, F. P. & Leben, L. 1940 *Phil. Trans.* A, **239**, 1.

Bowden, F. P., Moore, A. J. W. & Tabor, D. 1943 *J. Appl. Phys.* **14**, 80.

Bowden, F. P. & Ridler, K. E. W. 1936 *Proc. Roy. Soc.* A, **154**, 640.

Dacus, E. N., Coleman, F. & Roess, L. 1944 *J. Appl. Phys.* **15**, 813.

Finch, G. I. & Zahoorbux, Z. O. 1937 *Inst. Mech. Engrs, General Discussion on Lubrication*, **2**.

Lewis, C. 1940 *Proc. of the Special Summer Conferences on Friction and Surface Finish, Mass. Inst. Tech.*, Cambridge, Mass., 5, 6 and 7 June 1940, p. 212.

A cloud-chamber investigation of penetrating showers

By G. D. Rochester, *The Physical Laboratories, The University, Manchester*

(*Communicated by P. M. S. Blackett, F.R.S.—Received* 18 *February* 1946)

[Plates 7–9]

An account is given of a cloud-chamber investigation of penetrating showers. It is concluded that the cloud-chamber data are consistent with Jánossy's counter data and also with the view that penetrating showers consist of a small number of penetrating ionizing particles which are mainly mesons. There is no evidence to show that the showers are produced by processes other than those postulated in the theory of meson production proposed by Hamilton, Heitler & Peng, together with the addition suggested by Jánossy, i.e. the emission of small groups of mesons every few centimetres of lead. There are given several photographs of showers of penetrating particles which may be examples of the successive production of penetrating particles by an incident nucleon and its recoil particles. The spectrum of the penetrating particles seems to be approximately of the same form as the meson spectrum at sea-level.

Some 20 % of penetrating showers are accompanied, in the cloud chamber, by what appear to be electron cascades. It is shown that these showers cannot be due to knock-on electrons, to high-energy electron cascades penetrating the whole thickness of the absorber, or to decay electrons unless a meson of lifetime less than 10^{-10} sec. is postulated. It is suggested that these showers may be due to electrons or photons produced in processes which become important at very high energies, e.g. $> 10^{11}$ eV.

1. Introduction

The counter experiments of Wataghin and his co-workers (1940, 1941), and of Jánossy and his co-workers (1940, 1942), have established the existence at sea-level of showers differing from cascade and knock-on showers which are capable of penetrating 50 cm. of lead. These penetrating showers have been shown by Jánossy & Rochester (1943) to be produced almost equally by ionizing and non-ionizing particles. The showers show a transition effect which appears to be mass-proportional. With a layer of lead 5 cm. in thickness above the pentrating shower set approximately 50 % of the penetrating showers are created in the lead. The other 50 % come from the air surrounding the apparatus and are probably parts of extensive penetrating showers.

The recent theory of meson production proposed by Hamilton, Heitler & Peng (1943) seems to give a very plausible explanation of penetrating showers. According

---

[*Note added in proof* 10 *September* 1946.] Dr E. P. George of Birbeck College, London University, has kindly informed the author of an experiment he has recently carried out showing the occurrence of penetrating particles in bursts. In George's arrangement the bursts are recorded by an ionization chamber (surrounded by 7 cm. lead), placed above a counter set which selects penetrating showers. George concludes from the results that the bursts are produced in the 7 cm. lead by extensive (Auger) showers, and that the penetrating particles are produced at the same time in the ratio of one penetrating particle to thirty electrons and photons. The bursts connected with penetrating showers described in the present paper cannot be accounted for in a similar way for they occur below 37 cm. lead.

to this theory the showers consist mainly of charged mesons and neutral mesons (neutrettos), produced by protons and neutrons, in reactions of the following types:

$$\begin{aligned} P+P &\rightarrow N+P+Y^+, \\ N+P &\rightarrow P+P+Y^-, \\ P+Y^- &\rightarrow N+Y^0, \\ N+Y^+ &\rightarrow P+Y^0. \end{aligned}$$

The symbols P and N represent protons and neutrons, Y^+ and Y^- charged mesons, and Y^0 neutrettos. The charged mesons are of two types, one with a lifetime at rest of 10^{-6} sec., which has been observed experimentally, and the other a hypothetical particle with a lifetime of 10^{-8} sec. The neutretto is also a hypothetical particle since it has not been observed experimentally. In order that these reactions may occur, incident nucleons must have energies greater than 2×10^9 eV. Jánossy (1943) has shown that, particularly for heavy elements, one should expect several mesons emitted from the same nucleus. The calculated total number of mesons of all kinds, neutral as well as ionizing, with energies greater than 5×10^8 eV produced when a nucleon of energy greater than 10^{11} eV crosses an oxygen nucleus is 5 and for a lead nucleus is 18. To these numbers should be added a small number (<3) of recoil protons and neutrons (Peng 1944). These numbers are, of course, only average numbers and large fluctuations may occur. Penetrating showers are therefore expected to consist of *small* numbers of ionizing penetrating particles formed by a nucleon every few centimetres of lead. The production of small groups of mesons in succession will be referred to throughout this paper as plural meson production.*

The published photographs of penetrating showers fall into two main classes, those showing pairs of penetrating particles and those showing groups of penetrating particles which in every case appear to originate at one point. Typical of the first class are the photographs of Braddick & Hensby (1939), J. G. Wilson (1940), Leisegang (1940) and Seren (1942), taken near sea-level, and of Powell (1940), Hughes (1941), Herzog & Bostick (1941), taken at moderately high altitudes. Photographs of groups of particles are rarer than pairs, but good examples have been obtained by Jánossy, McCusker & Rochester (1941), Bose, Choudhuri & Sinha (1944), and R. P. Shutt (1946), at sea-level, and by Hazen (1943), Powell (1941), Daudin (1944), Bostick (1942), and Wollan (1941) at altitudes of from 6000 to 14,000 ft. above sea-level. A photograph reproduced by Euler & Heisenberg (1938), taken by Fussell, is also clearly a penetrating shower of the same type.

There is thus ample cloud-chamber evidence that small groups of penetrating particles do occur. No systematic cloud-chamber investigation of penetrating showers has, however, been undertaken as yet. The difficulty arises from the fact that, especially at sea-level, penetrating showers are very rare. Because of this the only efficient method of investigation is to use a cloud chamber controlled by a counter arrangement highly selective for penetrating showers. Counter control with an arrangement of this type results not only in a considerable saving of time and film but reveals features in penetrating showers which otherwise could not be recognized. In the investigation described in the present paper 95 % of the photographs are significant for penetrating showers. Had rigid selection not been used 85 % of the photographs would not have been recognized as penetrating showers. Rigid counter selection may, however, result in the loss of showers of certain types. Photographs have therefore also been taken with a less rigid selection; in this case the only criterion for a penetrating shower is the presence of two or more ionizing penetrating particles.

---

* I am indebted to Dr J. G. Wilson for this suggestion.

2. EXPERIMENTAL ARRANGEMENT

The experimental arrangement, illustrated diagrammatically in figure 1, consisted of a cloud chamber (C) operated by a set of fourteen counters arranged in sevenfold coincidence. The counters were placed in three trays with three sets of two counters in tray T and two sets of two counters in each of the trays M and B. Counters connected in parallel are shown in the diagram connected by lines. Three ionizing particles were required to set off the top tray and two to set off each

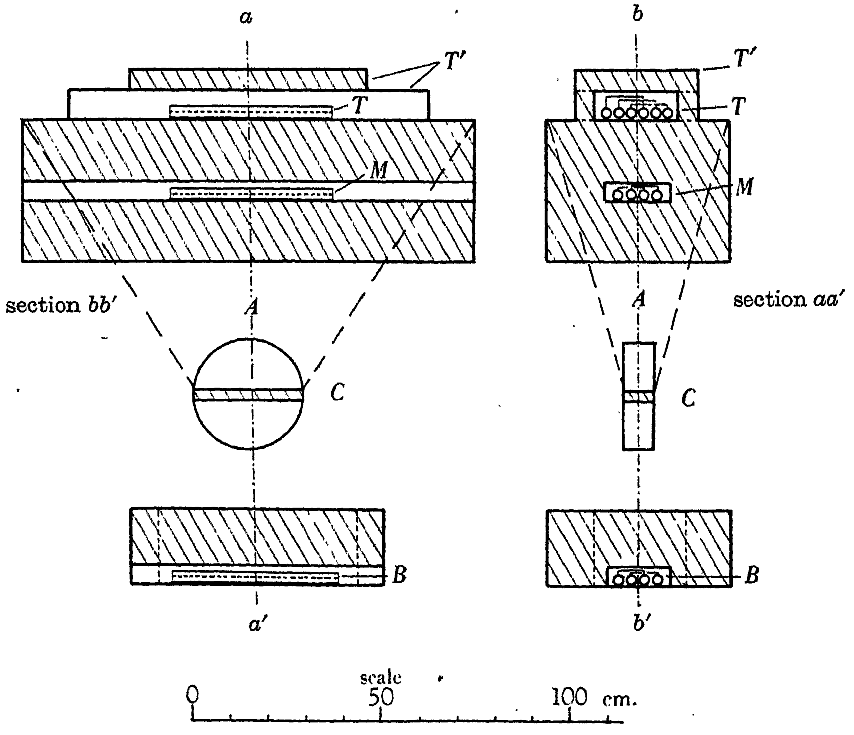
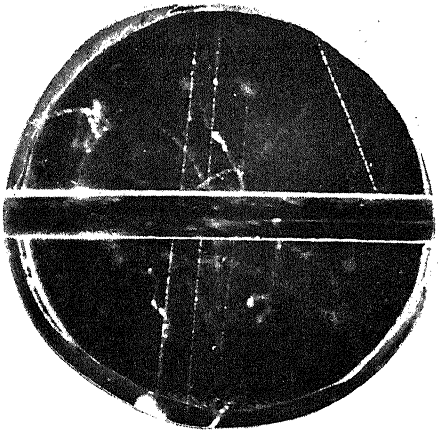
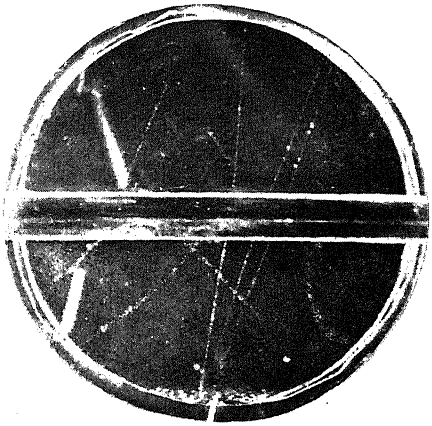
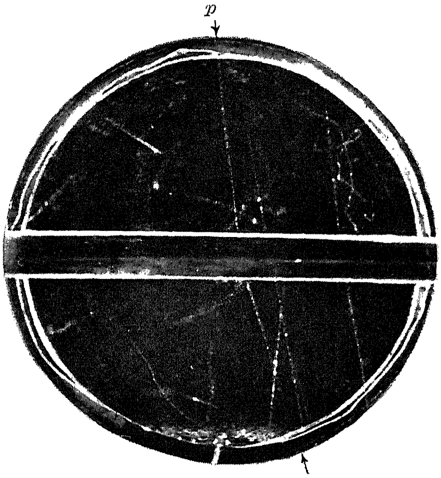


FIGURE 1. Experimental arrangement. T, M, B, counters; C, cloud chamber.

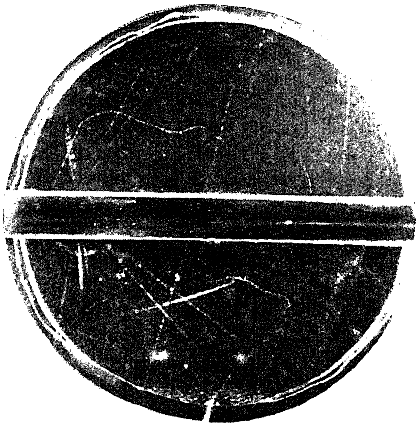


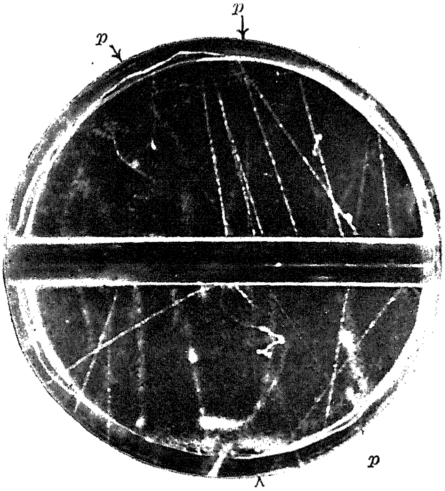
4



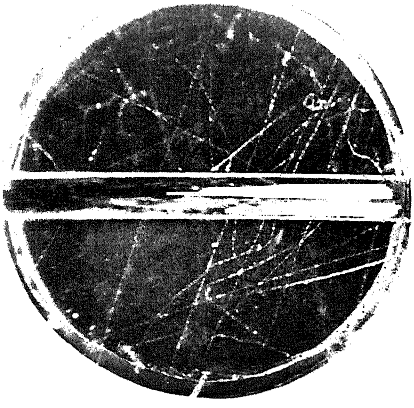


6

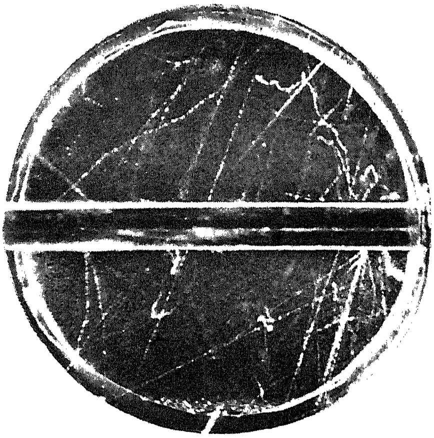




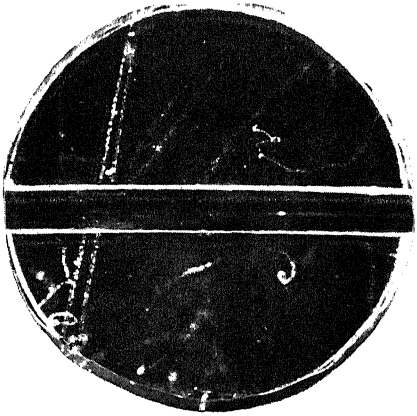
8



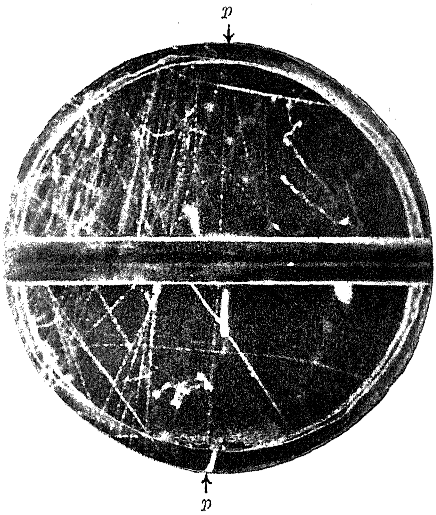
9



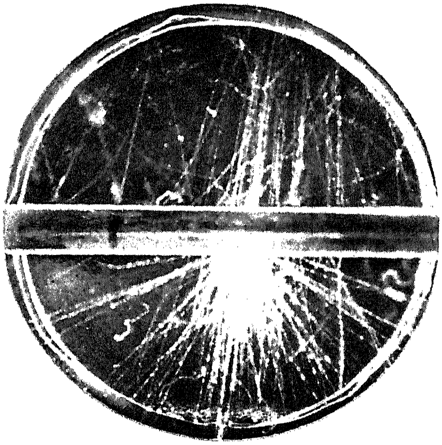
10



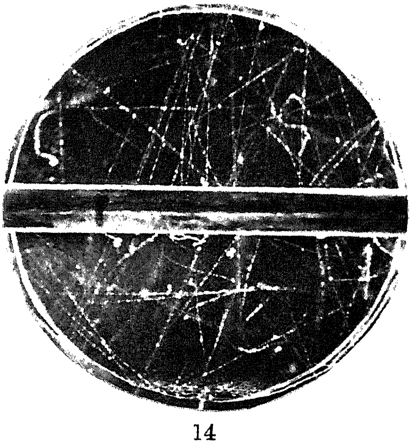
11



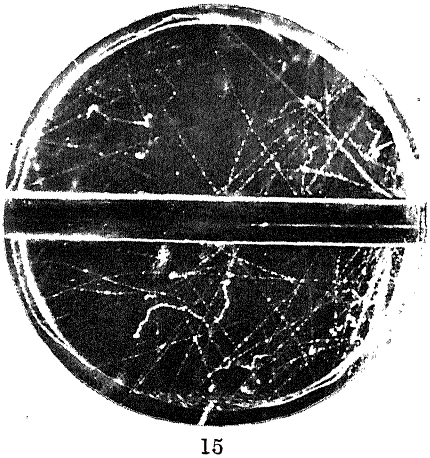
12



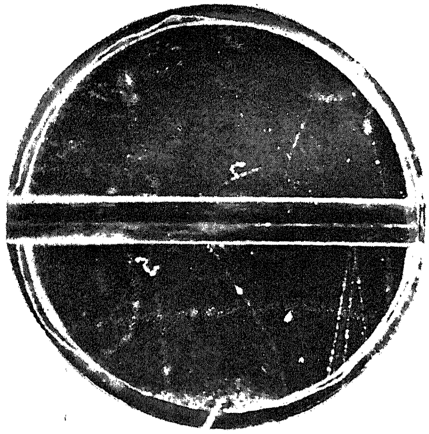
13



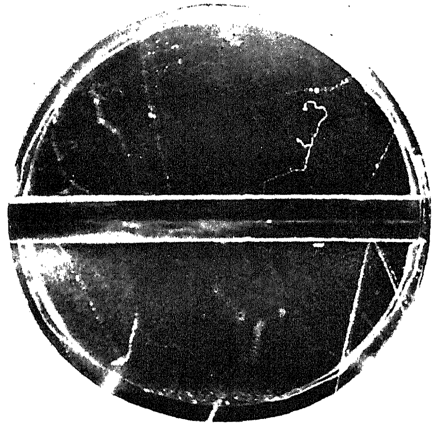
14



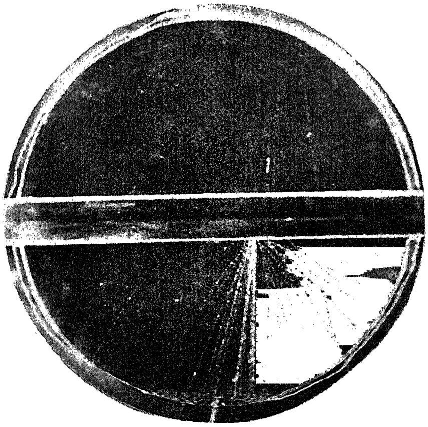
15



16



17



18

of the other trays. The counters were of the copper-in-glass type, 40 cm. long and 3 cm. internal diameter, and were filled with the usual mixture of 1·5 cm alcohol vapour and 10·0 cm. argon. Electron cascade showers from the air were eliminated by surrounding the middle and bottom trays with lead 15 cm. in thickness. The total vertical thickness of lead was 53 cm. A layer of lead 5 cm. in thickness was placed above and at the sides of the tray T in order to increase the rate of penetrating showers. Jánossy & Rochester (1944) showed that a sevenfold coincidence set with the counters disposed as described was highly selective for penetrating showers, and this conclusion was fully confirmed in the present experiment.

Before the cloud chamber was set into operation careful tests were made with the counter set alone to find the rate of sevenfold coincidences and the contribution of processes other than penetrating showers. The counting rate was $0\cdot026 \pm 0\cdot003$ counts per hr. It was shown by Jánossy (1942) that the main processes which simulate penetrating showers were knock-on showers and accidental coincidences. In the present experiment the rate of knock-on showers was estimated as 3×10^{-4} c.p.hr. and the rate of accidental coincidences as 6×10^{-4} c.p.hr. Thus the total contribution from knock-on showers and accidental coincidences was 9×10^{-4} c.p.hr.; one was therefore justified in assuming that not more than 5 % of the counts were due to processes other than penetrating showers.

The cloud chamber was placed between the middle and bottom trays. It was 30 cm. in diameter and had an illuminated depth of approximately 8 cm. Across the middle of the chamber was placed a lead plate 2·3 cm. in thickness; a particle which passed through this plate without multiplication and with small scattering was classified as 'penetrating' (see § 3). Illumination for the chamber was provided by two banks of 230 V 200 W tungsten coil-coil lamps flashed for approximately 0·2 sec. at 400 V a.c. The expansion mechanism was similar to that described by Blackett (1934).

The chamber was filled with a mixture of argon and oxygen in the proportion of 2:1 by volume to a pressure of 100 cm. Hg. The usual mixture of alcohol and water served to moisten the gas.

A voltage of 120 V was applied across the chamber and was reduced to approximately 10 V by a gas-filled relay slightly before and during the expansion.

Stereoscopic photographs were taken with a pair of matched Ross lenses of focal lengths 1·5 in. and apertures of 1·9, on Selo H.P. 3 film. The average rate of photographs was the same, within the standard deviation, as the rate of counts.

3. PENETRATING PARTICLES

In this section the evidence given by the photographs for the existence of penetrating particles in the showers is discussed.

(*a*) *The identification of penetrating particles.* A particle is classified as 'penetrating' if it can penetrate the 2·3 cm. lead plate and emerge as a single particle without appreciable scattering. The relevant data are given in tables 1A and 1B, and from these data it will be shown that there is only a small chance of confusing penetrating particles and electrons above 5×10^8 eV.

TABLE 1A. PROBABILITIES $\{P(N)\}$ OF LESS THAN THREE PARTICLES EMERGING FROM THE 2·3 CM. LEAD PLATE AS A FUNCTION OF THE ENERGY (E_e) OF AN INCIDENT ELECTRON

E_e (eV)	$\bar{N}_A$ (Arley)	$\bar{N}_B$ (Bhabha & Chakrabarty)	$P(N)$ (Pólya) %	$P(N)$ (Poisson) %	$\bar{\theta}$ (deg.)
(i)	(ii)	(iii)	(iv)	(v)	(vi)
$7\cdot4\times10^7$	0·8	0·5	90	95	20
$5\cdot5\times10^8$	10·5	8·0	33	0·18	2·5
$4\cdot0\times10^9$	64·0	50·0	7	3×10^{-23}	0·3

TABLE 1B. PROBABILITIES OF FINDING N PARTICLES BELOW THE 2·3 CM. LEAD PLATE FOR INCIDENT ELECTRONS OF ENERGY E_e

N	E_e (eV)	$7\cdot4\times10^7$	$5\cdot5\times10^8$	$4\cdot0\times10^9$
0	Pólya	0·651	0·133	0·015
	Poisson	0·449	3×10^{-5}	$1\cdot5\times10^{-28}$
1	Pólya	0·162	0·108	0·028
	Poisson	0·359	0·00028	$9\cdot6\times10^{-27}$
2	Pólya	0·075	0·078	0·023
	Poisson	0·144	0·0015	$3\cdot1\times10^{-25}$

In the table are given the average number of particles ($\bar{N}$) emerging from the lead plate when the incident particle is an electron of energy E_e. Data in column (ii) are taken from Arley (1938) and in column (iii) from the more accurate calculations of Bhabha & Chakrabarty (1943). As the differences between the two sets of data are not significant for present purposes, Arley's data will be used. In columns (iv) and (v) are given the percentage probabilities of finding less than three particles below the lead plate with Pólya and Poisson distributions. The detailed probabilities for $N=0$, 1 and 2, when the average is $\bar{N}_A$ (column (ii), table 1A), for different values of E_e are given in table 1B. The values for the Pólya distribution have been taken from Arley (1943), who has shown that this distribution gives a more accurate picture of the growth of an electron cascade shower than the Poisson distribution because the former assumes a growth analogous to biological growth, in which one generation succeeds another, whereas the latter assumes shower particles to be independent. The number N includes all particles, primary as well as secondary. In column (vi) are given values of the mean projected angle of scattering calculated from Williams's scattering formula (1939),

$$\overline{E\beta\theta} = (19\cdot5 - 3\cdot1 \log_{10} Z)^{\frac{1}{2}}\, 600 Ze(Nt)^{\frac{1}{2}},$$

where Z is the atomic number, e the electronic charge in e.s.u., N is the number of atoms per c.c., t is the thickness of the absorber in cm., $\beta=\mathrm{v}/c$. and E is the momentum of the particles (i.e. $300H\rho$ eV/c.). The value of $\overline{E\beta\theta}$ for fast particles traversing the lead plate is $1{\cdot}38 \times 10^9$ eV/c.deg. Thus a particle with an energy of 10^9 eV would be scattered on the average through 1·38° in the plane of the photograph.

It is concluded from table 1A that the identification of penetrating particles which are scattered through less than 3° is fairly certain above $5{\cdot}5 \times 10^8$ eV. The identification is, in fact, certainly better than would appear from table 1A, because many of the observed penetrating particles come into the cloud chamber singly, whereas an electron would be generally accompanied by other electrons. It is therefore considered that identification is reliable for more than 90 % of the penetrating particles.

Mesons with momenta below 2×10^8 eV/c. can be distinguished from electrons by the increase in ionization in passing through the lead plate (Rossi & Greison 1941).

(*b*) *Types of showers of penetrating particles taken with the sevenfold coincidence arrangement.* Some examples of showers of penetrating particles are given in figures 4 to 8, plate 7. In figure 4 three penetrating particles can be seen and a fourth particle which is stopped by the plate. Figures 5 and 6, plate 7 are good examples of showers of at least three penetrating particles. It will be noticed that, particularly in the case of figure 5, the tracks cannot be reprojected back to a single point, suggesting that they might have been formed in succession by a single proton or neutron and their recoil particles. An equally valid explanation is, however, that the penetrating particles have been formed at one point, but that the tracks do not reproject back to a point because of scattering in the lead above the chamber.

Other examples of showers of penetrating particles are shown in figures 7 and 8, plate 7. A single penetrating particle can be seen in figure 7, and to the left side an explosion has occurred in the plate. Figure 8 is an example of a shower of particles most of which are stopped in the plate, and two accompanying penetrating particles (marked '*p-p*') which can be seen faintly in the background. There are so many tracks that it is difficult to correlate tracks above the plate with those below. The two penetrating particles which can be definitely identified make such a large angle with each other that it is doubtful if they can have been created at one point.

(*c*) *Examples of showers of penetrating particles taken with a fivefold coincidence arrangement.* Although the sevenfold coincidence gives a rigid selection of penetrating showers, the selection might be so rigid that other interesting showers are excluded. The arrangement might, for example, have a strong bias against showers consisting of pairs of penetrating particles, a type observed frequently before. A coincidence arrangement of lower multiplicity would tend to remove this bias. For this reason a number of photographs have been taken with a fivefold coincidence arrangement which consisted of two counters out of each of the trays *T*, *M* and *B*, chosen such that they were directly above and below the chamber. The two counters in the top tray were connected in parallel. Estimates of the contributions of knock-on showers and accidental coincidences to the fivefold rate showed that nearly all the photographs were due to double knock-on showers. In this case, therefore, the only criterion which could be used to select a penetrating shower was the presence of more than one penetrating particle. Three examples of showers with this arrangement are shown in figures 9 to 11. Figure 9, plate 7 is a good example of a shower containing at least four penetrating particles, and it will be noticed that the tracks are almost parallel. This is the largest shower of penetrating particles obtained in the present experiment. It may well be an example of plural meson production. Figure 10, plate 8 shows two penetrating particles with an associated shower starting in the plate. Figure 11, plate 8 is one of the few examples we have obtained of a pair of penetrating particles. Unfortunately, the photograph is technically poor, for the tracks are broadened by field doubling and a slight shift occurs between the upper and lower parts of the chamber. Nevertheless, the most probable explanation of the photograph is that it represents a pair of penetrating particles starting in the lead immediately above the chamber.

(*d*) *The density of the showers.* Using the criteria outlined in § 3 (*a*), fifty photographs taken with the sevenfold arrangement have been examined for penetrating particles. The size distribution is given in table 2.

TABLE 2

number of penetrating particles per photograph (N)	number of photographs with N penetrating particles
0	32
1	10
2	3
3	5

The most striking feature of the distribution is the low density of the showers.

(*e*) *The scattering of the penetrating particles.* Detailed measurements have been made of the scattering, in the 2·3 cm. lead plate, of thirty-two penetrating particles selected from sixteen photographs taken with the sevenfold arrangement, and two photographs taken with the fivefold arrangement. The measurements were made with a microscope fitted with a goniometer eyepiece, and four independent settings were made on the tracks in the upper and lower parts of the chamber for each stereoscopic picture. The angles of scatter are therefore the mean of eight independent readings. The average setting error on good tracks was about 0·3°. The results are given in figure 2*a*. Since the main error was likely to be distortion in the cloud chamber a number of photographs of single mesons were taken with a replica hollow wooden box in place of the 2·3 cm. lead plate. The box was made of light wood 4 mm. in thickness, and the surfaces were thickly gilded to prevent the scattering of light. On the assumption that the gold was 0·02 mm. in thickness the total scattering was given by $\overline{E\beta\theta} = 1 \times 10^8$ eV/c.deg. Since at least 95 % of the mesons have momenta above $2{\cdot}5 \times 10^8$ eV/c., the scattering produced by the box can be neglected. The results of measurements on twenty-six tracks, given in figure 2*b*, shows that all the particles are scattered by less than ± 2°·5. It is clear that this scattering is spurious and is due to distortion in the cloud chamber. Thus because of the possibility of distortion no reliance was placed on scattering angles less than 2°. It should be added that the tracks with the lead plate in the chamber were uniformly better than those taken with the wooden box.

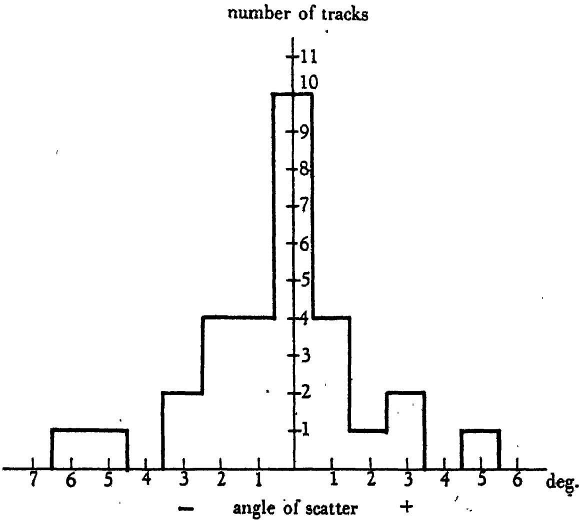


FIGURE 2*a*. Scattering of the penetrating particles in the 2·3 cm. lead plate.

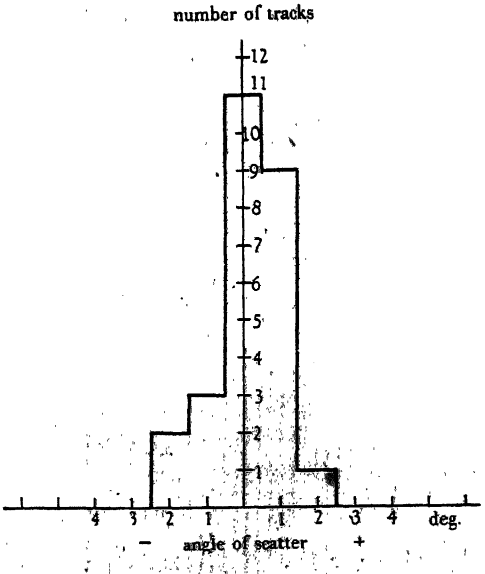


FIGURE 2*b*. Cloud-chamber distortion of tracks.

The distribution of scattering angles is given in table 3, from which it is seen that only two of the particles are scattered through more than 6°. A rough-energy spectrum has been obtained by using Williams's scattering formula (§ 3(*a*)), and the results are given in column (iv). The percentages of particles observed between these energy limits are given in column (iii), and these values may be compared with the actual percentages of mesons in cosmic radiation at sea-level, given in column (v), taken from the data of J. G. Wilson (1946). It is seen that there is significant agreement in spite of the large statistical fluctuations and the use of projected angles instead of actual angles of scatter.

TABLE 3. SCATTERING OF PENETRATING PARTICLES IN THE 2·3 CM. LEAD PLATE

range of $\bar{\theta}$ (deg.) (i)	number of tracks (ii)	percentage of total number $\bar{\theta} < 6°$ (iii)	energy limits from Williams's scattering formula (eV) (iv)	percentage of mesons in energy range given by (v) from Wilson's data (v)
0·0– 1·9	21	70	$7{\cdot}3 \times 10^8$–∞	86
2·0– 3·9	6	20	$3{\cdot}5 \times 10^8$–$7{\cdot}3 \times 10^8$	11
4·0– 5·9	3	10	$2{\cdot}3 \times 10^8$–$3{\cdot}5 \times 10^8$	3
6·0– 7·9	0	0	—	—
8·0– 9·9	0	0	—	—
10·0–11·9	1	—	—	—
18·0–19·9	1	—	—	—

Two of the penetrating particles are scattered through large angles, namely 10·2 and 18·0°. The scattering of the first of these particles can be fully accounted for in terms of multiple scattering. The scattering of the second particles may be multiple if the particle is an electron or 'large-angle' if the particle is a meson or a proton.

4. COMPARISON WITH JÁNOSSY'S COUNTER DATA (1942).

It is of interest to compare the present data with Jánossy's counter data both as regards the absolute rate of penetrating showers and the size distribution of the showers.

(*a*) *The absolute rate.* Jánossy found that the rate of penetrating showers was 1/12,000 of the rate of cosmic-ray particles penetrating 50 cm. of lead. The relevant data for the present experiment are collected in table 4. The threefold rate gives the total flux of cosmic-ray particles through the apparatus. Taking the sevenfold rate to be 0·030 c.p.hr., the ratio of the rate of penetrating showers to cosmic-ray particles which can penetrate 50 cm. lead is

$$0{\cdot}030:315 = 1:10{,}000.$$

This figure is so close to Jánossy's that it gives strong support to the conclusion that the same phenomenon is being observed in the two experiments.

TABLE 4

multiplicity	number of counters	counts	photographs	time (hr.)	rate per hr. ± S.D.
threefold	14	557	—	1·75	315 ± 13
sevenfold	14	71	—	2710·91	0·026 ± 0·003
sevenfold	14	—	50	1479·96	0·034 ± 0·005

(*b*) *The size distribution of the showers.* The size distribution of the showers has been obtained by counting the number of particles entering the top of the lead plate placed across the chamber. In order to exclude most electrons of radioactive origin, i.e. all below 2×10^6 eV, only particles leaving straight tracks have been counted. (The average projected angle of scattering of an electron of 5×10^6 eV, in a path 10 cm. long in the chamber is $\sim 4°$.) Jánossy's data refer to counters surrounded on all sides by 50 cm. lead, whereas the cloud chamber had no lead round its sides but only a relatively thin layer of absorber, mainly glass a few millimetres in thickness. For this reason, data have also been obtained for the number of particles entering the lead plate within an angle marked *A* in figure 1 which was well covered by the lead absorber above the chamber. These data are given in figures 3*a* and 3*b*. These data confirm Jánossy's conclusion that penetrating showers are a mixture of showers of different sizes, the larger part being showers of low density.

It will be noticed that 20–30 % of the photographs show no particles entering the top of the lead plate. These blank photographs cannot have been due to accidental coincidences, for the rate of accidental coincidences was not more than 2 % of the rate of penetrating showers. Thus it is unlikely that more than *one* blank photograph could be due to an accidental coincidence. It is, in fact, highly improbable that even one photograph was due to this cause for most of the accidentals were due to overlaps between pulses from sixfold coincidences and single counters. A sixfold coincidence would in general result only from a shower which would have several particles in the chamber. Further, none of the blank photographs can be explained by 'knock-on' showers; for the measured rate of this type of shower is exceedingly small.

It is concluded that none of the photographs with $N = 0$ are spurious. These photographs provide additional evidence that most penetrating showers are of low density. Reference to figure 1 will make it clear that a shower of low density could miss the lead plate in the chamber.

Since the illuminated part of the lead plate is approximately equal in area to one of Jánossy's *H* counters, it is possible to make a direct comparison of the probability of discharging one particular *H* counter with the probability of finding one or more particles entering the lead plate. The probability of discharging one *H* counter is given by the following expression:

$$P(1) = \frac{a_1 + 2a_2 + 3a_3 + \ldots + 8a_8}{8(a_0 + a_1 + a_2 + \ldots + a_8)},$$

where a_n represents the number of cases in which n counters are discharged, and n takes the values $0\rightarrow 8$. The value of $P(1)$, calculated from Jánossy's data after making allowance for the fact that most of the records $n = 0, 1$ are due to 'knock-on' showers, is

$$P(1) = 0{\cdot}4 \pm 0{\cdot}1.$$

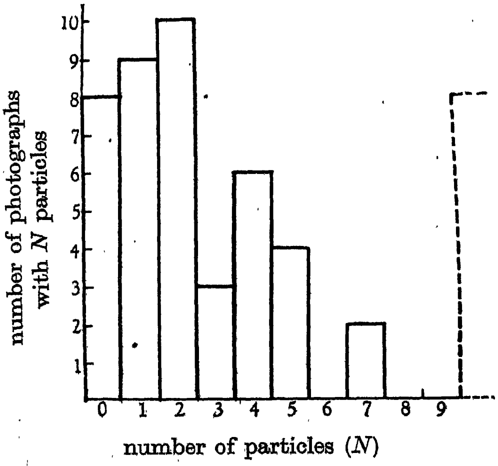


FIGURE 3*a*. Size distribution of the showers for particles entering the lead plate from any direction.

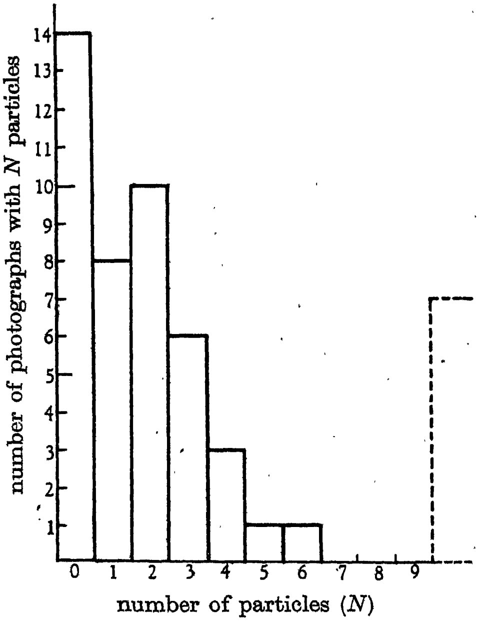


FIGURE 3*b*. Size distribution of the showers for particles entering the lead plate within the solid angle A (figure 1).

The corresponding value from cloud-chamber data is 0·7 ± 0·1. The two values are sufficiently close to show that both sets of data indicate showers of low density. The small difference between the two values may be due to the following factors. (1) The areas of the illuminated part of the lead plate and a single H counter are almost the same, but the dimensions are quite different, being 28 × 8 cm.^{2} for the lead plate and 3·5 × 60 cm.^{2} for an H counter. A change in geometrical form might have a large effect on the value of $P(1)$. (2) The value of $P(1)$ refers to any counter out of the H set whereas the cloud chamber is in a position corresponding to one of the middle counters.

5. LARGE SHOWERS

(*a*) *Frequency and size distribution.* Although most penetrating showers consist of showers of low density there occur also a number of showers of surprisingly high density. Data given in table 5 indicate the number of photographs on which there are not less than ten particles within the solid angle A (figure 1) above the lead plate. Examples of these large showers are given in figures 12 to 15, plate 8.

TABLE 5

number of particles (N) (i)	number of photographs with not less than N particles (ii)	% number of photographs with not less than N particles (iii)	theoretical number of showers with not less than N particles (knock-on theory) (iv)
10	7	14	0·0950
20	4	8	0·0180
30	3	6	0·0060
40	3	6	0·0010
50	2	4	0·0007

Figures 12 and 13 are examples of the very large showers of some forty and fifty particles respectively. Figure 12 is of especial interest because it is a large shower which definitely shows an associated penetrating particle (marked 'p-p' in the figure). Approximately half of the large showers seem to have associated penetrating particles, but a number of the showers are so dense that it is difficult to say whether penetrating particles are present or not.

Both of the largest showers observed are of the form of the shower shown in figure 13.

(*b*) *The origin of the large showers.* These large showers occur below 40 cm. of lead and consist mainly of particles which are stopped in the 2·3 cm. lead plate. Moreover, they must have at least three particles at the tray T, two particles at the tray M, ten particles in the cloud chamber, and two at the tray B. It is therefore clear that part of the showers must be very penetrating. It can be assumed that the penetrating part consists of ionizing particles such as those observed on other photographs of penetrating showers (§ 3). In this case the showers would be 'soft' secondary radiation accompanying the penetrating particles. This description would commend itself if the 'soft' secondary particles could be ascribed to a knock-on process in which a meson knocks on an electron in the lead close above the cloud chamber. The electron then produces the cascade shower observed in the chamber. The meson should accompany the shower, but it may not be observed if the shower is very dense or if the core of the shower lies to the edge of the illuminated part of the chamber. The probability of getting a shower of a certain size is a function of the probability of a meson of given energy producing a knock-on electron and the number of mesons with this energy. Lovell (1939) has calculated this probability, and from his results the number of showers to be expected in the cloud chamber has been estimated. Each penetrating shower has been assumed to contain ten mesons. The results are given in column (iv) of table 5. The observed values are given in column (ii), and a comparison of the two sets of figures shows clearly that the large showers are not knock-on showers.

The second possibility is that the showers are high-energy electron cascades which have traversed the whole of the lead absorber. Jánossy has estimated that a cascade shower which can penetrate 50 cm. of lead would require an electron of energy 10^{19} eV. Assuming an incident spectrum of the form

$$\int S(E)\,\delta E = H/E_0^{\gamma},$$

where $H = 6 \times 10^9$ hr./m.^{2}, $\gamma = 1{\cdot}50$ and E_0 is the energy of the electron measured in units of the critical energy, the rate of electrons with energies 10^{19} eV is estimated at 2×10^{-7}/m.^{2}/hr. Thus the rate of high-energy electrons on the absorber T' (area $0{\cdot}1$ m.^{2}) above the cloud chamber would be 2×20^{-8}/hr. Taking the total time of the experiment to be $50 \times 35 = 1750$ hr., the expected number of large showers would be 3×10^{-5}. This number is again so small that none of the observed photographs is likely to be a large cascade shower which has penetrated the lead. This estimate would be invalidated if there were any breakdown in the fundamental processes of radiation loss and pair production at very high energies.

A third possibility is that the showers are large meson showers. This is unlikely for the following reasons: (1) many of the showers have the typical form of electron cascade showers, (2) most of the particles in the showers have ranges less than the thickness of the lead plate; if these particles were all mesons they would have ion densities appreciably greater than is observed; and (3) very few penetrating particles are observed in the large showers.

A fourth possibility is that the showers are electron cascades resulting from decay electrons produced in the lead above the cloud chamber. The probability of getting a shower of a certain size is a function of the probability of a meson of certain energy producing a decay electron and the number of mesons with this energy. An electron which can produce ten or more particles in a layer of lead a few centimetres in thickness must have an energy greater than 10^9 eV. Thus the least energy of the meson which can produce such a shower is approximately 2×10^9 eV. It is clear that the chance of the decay of an ordinary meson of energy greater than 2×10^9 eV in the lead absorber is negligibly small. It is worth examining, however, if the showers could result from the decay of mesons of lifetime 10^{-8} sec. or much less.

The probability of a meson of lifetime τ_0 decaying in a distance D is given by

$$p = \frac{D(1-\beta^2)^{\frac{1}{2}}}{\tau_0 c\beta} = \frac{D\mu}{\tau_0 E},$$

where μ is the mass of the meson at rest and E is its momentum. Let the number of mesons with momenta between E and $E+\delta E$ be $NS(E)\,\delta E$, where N is the total number of mesons above 2×10^9 eV. Thus the number (n') of large showers will be given by the formula

$$n' = \frac{D\mu N}{\tau_0} \int_{2\times 10^9}^{\infty} \frac{S(E)\,\delta E}{E}.$$

Assume that the mesons have the same energy spectrum and the same mass as the ordinary mesons found at sea-level. Assume further that each penetrating shower contains on the average ten mesons of all kinds; then N cannot be greater than 100. The expression under the integral sign can be evaluated numerically, and taking $D = 5$ cm. the total number of large showers is found to be 0·02. This number is to be set against the observed number of seven large showers. It is clear that the electron cascades cannot be due to the decay of a meson of lifetime 10^{-8} sec.; it would, in fact, require a lifetime less than 10^{-10} sec. to account for the showers in this way. There is, at present, no theoretical foundation for assuming the existence of such a meson.

Hamilton, Heitler & Peng have suggested that new processes are to be expected for energies greater than 10^{11} eV; it is therefore possible that the large showers are produced by the interaction of particles of very high energy with matter. Extrapolation of the ordinary meson spectrum to high energies shows that enough particles of very high energy fall on the absorber T' to account for the showers.

6. SHOWERS STARTING IN THE LEAD PLATE

Several examples of showers starting in the lead plate have been observed. Examples are given in figure 7, plate 7, and figures 16 to 18, plate 9. The shower shown in figure 16 seems to have been initiated by a neutral particle, but the shower of figure 17 may have been initiated by either a neutral or an ionizing particle. Two of the particles of the latter shower are heavily ionizing suggesting that they are slow protons or slow mesons. Another example of a shower is shown in figure 18. It is suggested that this shower might be a cascade shower, since it has the form of a cascade and a group of particles enter the top of the plate near the origin of the shower.

Hazen (1944) has observed many similar showers and he has concluded that they are due almost equally to high-energy neutrons and protons. The low-energy stars, which have been observed very frequently, are due mainly to neutrons. Particles produced in low-energy stars are observed to be distributed in all directions whereas the particles from 'disintegrations' produced by high-energy particles are largely collimated in the direction of the incident particle. The collimated stars are presumably penetrating showers starting in the lead plate. According to the theory of Hamilton, Heitler & Peng they would be expected to be produced equally by neutrons and protons.

On many of the photographs heavily ionizing particles are observed, e.g. figures 7, 9, 12, 15 and 17. Seventeen heavily ionizing particles have been observed in the fifty photographs.

I wish to thank Professor P. M. S. Blackett, F.R.S., for the interest he has shown in this work and for useful comment on the paper. I wish also to express my thanks to Dr L. Jánossy for the benefit of many valuable discussions and much help during the course of this work. I am indebted to Mr D. Broadbent, B.Sc., for help in running the cloud chamber during the latter stages of the work.

REFERENCES

Arley, N. 1938 *Proc. Roy. Soc.* A, **168**, 519.
Arley, N. 1943 *On the theory of stochastic processes.* Copenhagen, Denmark: G. E. C. Forlag.
Bhabha, H. J. & Chakrabarty, S. K. 1943 *Proc. Roy. Soc.* A, **181**, 267.
Blackett, P. M. S. 1934 *Proc. Roy. Soc.* A, **146**, 281.
Bose, D. M., Choudhuri, B. & Sinha, M. 1944 *Phys. Rev.* **65**, 341.
Bostick, W. H. 1942 *Phys. Rev.* **61**, 557.
Braddick, H. J. J. & Hensby, G. S. 1939 *Nature*, **144**, 1012.
Daudin, J. 1944 *Ann. Phys., Paris*, (11), **19**, 110.
Euler, H. & Heisenberg, W. 1938 *Ergebn. Exakt. Naturw.* **17**, 1.
Hamilton, J., Heitler, W. & Peng, H. W. 1943 *Phys. Rev.* **64**, 78.
Hazen, W. E. 1943 *Phys. Rev.* **64**, 257.
Hazen, W. E. 1944 *Phys. Rev.* **65**, 67.
Herzog, G. & Bostick, W. H. 1941 *Phys. Rev.* **59**, 122.
Hughes, D. J. 1941 *Phys. Rev.* **60**, 414.
Jánossy, L. & Ingleby, P. 1940 *Nature*, **145**, 511.
Jánossy, L. 1942 *Proc. Roy. Soc.* A, **179**, 361.
Jánossy, L. 1943 *Phys. Rev.* **64**, 345.
Jánossy, L., McCusker, C. B. A. & Rochester, G. D. 1941 *Nature*, **148**, 660.
Jánossy, L. & Rochester, G. D. 1943 *Proc. Roy. Soc.* A, **182**, 180.
Jánossy, L. & Rochester, G. D. 1944 *Proc. Roy. Soc.* A, **183**, 181.
Leisegang, S. 1940 *Z. Phys.* **116**, 515.
Lovell, A. C. B. 1939 *Proc. Roy. Soc.* A, **172**, 568.
Peng, H. W. 1944 *Proc. R. Irish Acad.* A, **49**, 245.
Powell, W. 1940 *Phys. Rev.* **58**, 474.
Powell, W. 1941 *Phys. Rev.* **60**, 413.
Rossi, B. & Greisen, K. 1941 *Rev. Mod. Phys.* **13**, 240.
Seren, L. 1942 *Phys. Rev.* **62**, 204.

Shutt, R. P, 1946 *Phys. Rev.* **69**, 261.
Wataghin, G., de Souza Santos, M. D. & Pompeia, P. A. 1940 *Phys. Rev.* **57**, 61.
Wataghin, G., de Souza Santos, M. D. & Pompeia, P. A. 1941 *Phys. Rev.* **59**, 902.
Williams, E. J. 1939 *Proc. Roy. Soc.* A, **169**, 531.
Wilson, J. G. 1940 *Proc. Roy. Soc.* A, **174**, 73.
Wilson, J. G. 1946 *Nature* (in the Press).
Wollan, E. O. 1941 *Phys. Rev.* **60**, 532.

DESCRIPTION OF PLATES 7–9

S, sevenfold; F, fivefold.

PLATE 7

FIGURE 4. A penetrating shower with at least three penetrating particles and a fourth particle which is stopped in the plate. (S 5.)

FIGURES 5 and 6. Photographs of penetrating showers containing at least three penetrating particles which do not appear to originate at a single point. These showers may be examples of plural meson production. One of the penetrating particles of figure 6 is scattered through 10·2°. (S 28 and S 48.)

FIGURE 7. A penetrating shower with at least one penetrating particle and an explosion in the plate. One of the particles emerging from the plate is so heavily ionizing that it is probably multiply-charged. (S 31.)

FIGURE 8. A penetrating shower in which most of the particles in the cloud chamber stop in the lead plate. The shower contains two penetrating particles marked '*p-p*' which may be seen faintly in the background. (S 47.)

FIGURE 9. A penetrating shower containing at least four penetrating particles with almost parallel tracks. This photograph seems to indicate plural meson production. (F 9.)

PLATE 8

FIGURE 10. A penetrating shower containing at least two penetrating particles and an associated electron-cascade shower which starts in the lead plate. (F 17.)

FIGURE 11. A pair of penetrating particles starting in the lead immediately above the cloud chamber. The photograph is technically poor, for there is field separation of the positive and negative ions and a slight shift between tracks in the upper and the lower parts of the chamber. (F 7.)

FIGURE 12. A large shower of some 40 particles associated with a penetrating shower. One penetrating particle, marked '*p-p*', can be clearly distinguished. (S 50.)

FIGURE 13. A large shower of at least fifty particles in the cloud chamber. It is thought that this shower is a large electron-cascade shower produced in the lead immediately above the cloud chamber. (S 45.)

FIGURES 14 and 15. Examples of large showers which are associated with penetrating showers. (S 49 and S 27.)

PLATE 9

FIGURE 16. An example of a shower which starts in the lead plate. (S 37.)

FIGURE 17. Another example of a shower starting in the lead plate. This shower seems to be an explosion which contains heavy particles. (S 21.)

FIGURE 18. An electron-cascade shower starting in the lead plate. (S 29.)

Stability of viscous flow between rotating cylinders. II. Cylinders rotating in opposite directions

By D. Meksyn, D.Sc.

(*Communicated by Professor G. Temple, F.R.S.—Received* 28 *February* 1946)

1. Introduction and summary

In a previous paper (Meksyn 1946; it will be referred to as Part I) the problem of stability of viscous fluid between coaxial cylinders, rotating in the same direction, was solved by expanding the integrals in inverse powers of a large parameter.

The question was first considered theoretically and experimentally by Taylor (1923), the problem being solved by expanding the integrals in orthogonal Bessel functions.

The aim of the present paper is to extend the solution to the case when the cylinders rotate in opposite directions.

The important difference between the two cases in its mathematical aspect consists of the following.

It is necessary to find the asymptotic integrals of a certain linear differential equation.

In the case when the cylinders rotate in opposite directions, these integrals become infinite within the range under consideration; namely, approximately at the point where the mean velocity of rotation is equal to zero, and the asymptotic expansions change their form in passing through this point.

It is, therefore, necessary to find the law of transformation of these integrals; that requires a rather extensive mathematical investigation.

For the sake of convenience the work is divided into two separate parts, hydrodynamical (Part II) and mathematical (Part III).

In the present paper (Part II) the transformations are only quoted, whereas in the mathematical part the detailed solution of the equations is developed.

The main results obtained are as follows.

The spacings of the vortices and the critical angular speed Ω_1 of the inner cylinder are found from the equations

$$\lambda d_1 = \frac{\pi}{1\cdot546}, \quad \frac{4\cdot05^3}{d_1} = \frac{8\Omega_1^2}{\nu^2\lambda^4}\frac{(1-\mu R_2^2/R_1^2)^2}{(R_2^2/R_1^2-1)^2 R_0},$$

where $d_1 = R_0 - R_1$; R_1 and R_2 ($R_2 > R_1$) are the radii of the cylinders, R_0 (2·10) is the radius where the mean velocity V vanishes, and $\mu = \Omega_2/\Omega_1$; $2\pi/\lambda$ is the wave-length parallel to the axis of the cylinder.

The critical velocities were evaluated and found in good agreement with Taylor's experimental results.

The agreement, however, greatly depends on the accuracy of measurement of R_1 when the distance between the cylinders is small and $-\mu$ is large.

2. EQUATIONS OF MOTION

For the sake of convenience certain results are given here which will be needed below; the details are found in Part I.

The radii of the two cylinders are denoted by R_1 and R_2 respectively ($R_2 > R_1$).

If r is the distance of a point from the axis, then the mean velocity V, in steady motion, is

$$V = Ar + \frac{B}{r}, \quad A = \frac{\Omega_1(1-\mu R_2^2/R_1^2)}{1-R_2^2/R_1^2}, \quad B = \frac{R_1^2\Omega_1(1-\mu)}{1-R_1^2/R_2^2}, \tag{2·1}$$

where Ω_1 and Ω_2 are the angular velocities of the two cylinders and $\mu = \Omega_2/\Omega_1$.

The velocities of the disturbed motion are u, $V+v$ and w, where u is along r, $V+v$ perpendicular to the meridian plane and w the component along the axis.

It is assumed that

$$u = u_0 \cos \lambda z\, e^{\sigma t}, \quad v = v_0 \cos \lambda z\, e^{\sigma t}, \quad w = w_0 \sin \lambda z\, e^{\sigma t}, \tag{2·2}$$

and it can be shown (Part I) that the solution is reduced to the integration of the following differential equation for v_0:

$$\begin{aligned}\frac{d^6v}{dr^6} + \frac{3}{r}\frac{d^5v}{dr^5} - \left(\frac{6}{r^2} + 2\lambda_1^2 + \lambda^2\right)\frac{d^4v}{dr^4} + \left(\frac{12}{r^3} - \frac{4\lambda_1^2 + 2\lambda^2}{r}\right)\frac{d^3v}{dr^3} \\ + \left(-\frac{27}{r^4} + \frac{6\lambda_1^2 + 3\lambda^2}{r^2} + \lambda_1^4 + 2\lambda^2\lambda_1^2\right)\frac{d^2v}{dr^2} + \left(\frac{45}{r^5} - \frac{6\lambda_1^2 + 3\lambda^2}{r^3} + \frac{\lambda_1^4 + 2\lambda^2\lambda_1^2}{r}\right)\frac{dv}{dr} \\ + \left[-\frac{45}{r^6} + \frac{6\lambda_1^2 + 3\lambda^2}{r^4} - \frac{\lambda_1^4 + 2\lambda^2\lambda_1^2}{r^2} - \lambda^2\lambda_1^4 - \frac{4\lambda^2}{\nu^2}\left(A^2 + \frac{AB}{r^2}\right)\right]v = 0,\end{aligned} \tag{2·3}$$

where for the sake of brevity the subscript of v was dropped; λ is the large parameter, and

$$\lambda_1^2 = \lambda^2 + \frac{\sigma}{\nu}. \tag{2·4}$$

The boundary conditions are (Part I)

$$v_0 = 0, \quad \frac{d^2v_0}{dr^2} = 0, \quad \frac{d^3v_0}{dr^3} - \lambda_1^2\frac{dv_0}{dr} = 0, \tag{2·5}$$

at $r = R_1$ and $r = R_2$.

To solve (2·3) assume

$$v = Ce^{\xi}, \tag{2·6}$$

where ξ is a function of r, and it contains the large parameter λ; and C is a slowly varying function of r.

Substituting (2·6) in (2·3), equating to zero the two terms of the highest order respectively, and integrating, then ξ' and C are given by

$$\left.\begin{aligned} f(\xi') \equiv \xi'^6 - (2\lambda_1^2+\lambda^2)\xi'^4 + (\lambda_1^4+2\lambda^2\lambda_1^2)\xi'^2 - \lambda^2\lambda_1^4 - \frac{4\lambda^2}{\nu^2}\left(A^2+\frac{AB}{r^2}\right) &= 0, \\ C = \frac{\text{const.}}{r^{\frac{1}{2}}[\partial f(\xi')/\partial \xi']^{\frac{1}{2}}}, & \end{aligned}\right\} \tag{2·7}$$

where primes denote differentiation with respect to r.

To obtain ξ integrate ξ' with respect to r, whence

$$\xi = \int_{r_0}^{r} \xi' dr, \tag{2·8}$$

where $r = r_0$ corresponds to the critical point of the asymptotic expansions, namely, where $\partial f(\xi')/\partial \xi'$ vanishes, i.e. $r = r_0$ is the root of the equation (Part III, § 4)

$$\lambda^2\lambda_1^4 + \frac{4\lambda^2}{\nu^2}\left(A^2+\frac{AB}{r^2}\right) = 0. \tag{2·9}$$

The solution of the equation (2·3) can now be completed, which *will only be given for the case when* $\lambda_1 = \lambda$, i.e. when $\sigma = 0$.

Let $r = R_0$ be the radius where the velocity V vanishes; thus from (2·1) it is found that

$$R_0 = \frac{(1-\mu)^{\frac{1}{2}} R_2}{(1-\mu R_2^2/R_1^2)^{\frac{1}{2}}}. \tag{2·10}$$

Put

$$r = R_0 + x, \tag{2·11}$$

then from (2·1)

$$A^2 + \frac{AB}{r^2} = \frac{2\Omega_1^2(1-\mu R_2^2/R_1^2)^2}{(R_2^2/R_1^2-1)^2 R_0} x, \tag{2·12}$$

where terms of order $(x/R_0)^2$ and higher were disregarded.

Denoting for brevity

$$h = \frac{8\Omega_1^2}{\nu^2\lambda^4} \frac{(1-\mu R_2^2/R_1^2)^2}{(R_2^2/R_1^2-1)^2 R_0}, \tag{2·13}$$

it is found that ξ' is equal to (2·7)

$$\xi' = \pm \lambda[1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}}, \tag{2·14}$$

where ω is any root of

$$\omega^3 = 1, \tag{2·15}$$

whence

$$\xi = \pm \lambda \int_{x_0}^{x} [1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}} dx, \tag{2·16}$$

where x_0 is the root of

$$1 + hx = 0. \tag{2·17}$$

ξ is evaluated for the case when $\omega = 1$, and x is negative, since only this expression will be needed in our computations.

Changing the notation, let

$$L = \int_{x_0}^{x} [(-hx)^{\frac{1}{3}} - 1]^{\frac{1}{2}}\, dx. \tag{2·18}$$

Putting
$$(-hx)^{\frac{1}{3}} - 1 = q^2, \tag{2·19}$$

then
$$L = -\frac{6}{h}[(-hx)^{\frac{1}{3}} - 1]^{\frac{3}{2}} \left\{\frac{(-hx)^{\frac{2}{3}}}{7} + \tfrac{4}{35}(-hx)^{\frac{1}{3}} + \tfrac{8}{105}\right\}, \tag{2·20}$$

whence the corresponding values of ξ are

$$\xi = \pm \lambda i L. \tag{2·21}$$

It is worth while noticing that the expression for L (2·20) is valid if $\omega \neq 1$, provided that a suitable cube root is taken; care, however, is needed to get the correct argument of L.

3. INTEGRALS AND THEIR TRANSFORMATIONS

Now construct the six integrals for positive and negative values of x respectively.

First find six particular integrals for positive and negative values of x. A certain set of solutions is assumed for positive x, and their transformations are given for negative x.

Let (2·15)
$$\omega_1 = e^{\frac{2}{3}\pi i}, \quad \omega_2 = e^{\frac{4}{3}\pi i}, \quad \omega_3 = e^{2\pi i}, \tag{3·1}$$

and denote
$$x = x_1, \quad \text{when} \quad r = R_1, \qquad x = x_2, \quad \text{when} \quad r = R_2. \tag{3·2}$$

The particular integrals are all of the form (2·6), (2·7)

$$\frac{e^{\xi}}{r^{\frac{1}{2}}[\partial f(\xi')/\partial \xi']^{\frac{1}{2}}}, \tag{3·3}$$

where ξ is found from (2·16); for the case when $\omega = 1$ the explicit expression of ξ is given by (2·20).

It is easily obtained from (2·7) that

$$\frac{\partial f}{\partial \xi'} = 6\xi'[\xi'^2 - \lambda^2]^2, \tag{3·4}$$

whence the assumed system of integrals for positive x is as follows (Part III, (9·8)):

$$\left.\begin{aligned}
Y_1 &= \frac{e^{-\frac{2}{3}\pi i}\, e^{\xi_1}}{r^{\frac{1}{2}}(\xi_1')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, & Y_2 &= \frac{e^{\frac{2}{3}\pi i}\, e^{\xi_2}}{r^{\frac{1}{2}}(\xi_2')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, \\
Y_3 &= \frac{e^{\xi_3}}{r^{\frac{1}{2}}(\xi_3')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, & Y_4 &= \frac{-e^{\frac{1}{3}\pi i}\, e^{\xi_4}}{r^{\frac{1}{2}}(\xi_4')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, \\
Y_5 &= \frac{-e^{\frac{2}{3}\pi i}\, e^{\xi_5}}{r^{\frac{1}{2}}(\xi_5')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, & Y_6 &= \frac{e^{\xi_6}}{r^{\frac{1}{2}}(\xi_6')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}.
\end{aligned}\right\} \tag{3·5}$$

The particular constants in the integrals were taken for theoretical reasons; namely, to simplify the laws of transformation in passing through the critical point.

The integrals for negative values of x are denoted by $Z_1, Z_2, \ldots, Z_6$, and are obtained from $Y_1, Y_2, \ldots, Y_6$ by substituting instead of x the quantity

$$x = (-x)\, e^{\pi i}. \tag{3·6}$$

For instance,
$$Z_1 = \frac{-e^{\xi_1}}{r^{\frac{1}{2}}(\xi_1')^{\frac{1}{2}}(-hx)^{\frac{1}{3}}}, \quad Z_4 = \frac{-e^{\xi_4}}{r^{\frac{1}{2}}(\xi_4')^{\frac{1}{2}}(-hx)^{\frac{1}{3}}}, \tag{3·7}$$

where the corresponding arguments of ω, ξ' and ξ are given in tables 1 and 2.

TABLE 1. $x>0$, $\arg x = 0$, $\xi = \lambda\int_{x_0}^{x} [1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}}\, dx$, $\arg\xi' = \arg\xi$.

Y_n	ξ_n	ω	$\arg\xi_n$	$\xi_n'(x_2) = [1+\omega(hx_2)^{\frac{1}{3}}]^{\frac{1}{2}}$	
Y_1	ξ_1	$e^{\frac{2}{3}\pi i}$	$\pi/3$	a	∞
Y_2	ξ_2	$e^{\frac{4}{3}\pi i}$	$2\pi/3$	$-\bar{a}$	0
Y_3	ξ_3	$e^{2\pi i}$	π	$-a_0$	0
Y_4	ξ_4	$e^{\frac{8}{3}\pi i}$	$4\pi/3$	$-a$	0
Y_5	ξ_5	$e^{\frac{10}{3}\pi i}$	$5\pi/3$	$\bar{a}$	∞
Y_6	ξ_6	1	0	a_0	∞

TABLE 2. $x<0$, $\arg x = \pi$, $\xi = \lambda\int_{x_0}^{x} [1+e^{\frac{1}{3}\pi i}\omega(-hx)^{\frac{1}{3}}]^{\frac{1}{2}}\, dx$, $\arg\xi = \arg\xi' + \pi$.

Z_n	ξ_n	ω	$\arg\xi_n$	$\xi_n'(x_1)$ $= [1+e^{\frac{1}{3}\pi i}\omega(-hx_1)^{\frac{1}{3}}]^{\frac{1}{2}}$	
Z_1	ξ_1	$e^{\frac{2}{3}\pi i}$	$3\pi/2$	b_0	p
Z_2	ξ_2	$e^{\frac{4}{3}\pi i}$	$11\pi/6$	b	∞
Z_3	ξ_3	$e^{2\pi i}$	$13\pi/6$	$\bar{b}$	∞
Z_4	ξ_4	$e^{\frac{8}{3}\pi i}$	$5\pi/2$	$-b_0$	p
Z_5	ξ_5	$e^{\frac{10}{3}\pi i}$	$17\pi/6$	$-b$	0
Z_6	ξ_6	1	$7\pi/6$	$-\bar{b}$	0

The letters a and b are introduced for the sake of brevity. The arguments of ξ' and ξ are evaluated for $|hx|$ infinite, except in the case of Z_1 and Z_4, where the arguments of ξ' and ξ are valid for finite (hx) provided that $|hx|>1$; in the last column it is indicated whether the integrals tend to infinity, to zero, or are periodic for large values of $|hx|$.

The system of integrals selected as a solution of (2·3) and their transformations, are given in table 3.

TABLE 3

	$x>0$, $\arg x = 0$		$x<0$, $\arg x = \pi$	
v_1	Y_1	∞	$-Z_1$	p
v_2	Y_2	0	$-Z_2+Z_1$	∞, p
v_3	Y_2+Y_3	0, 0	Z_3+Z_1	∞, p
v_4	Y_4	0	Z_4-Z_3	p, ∞
v_5	Y_5-Y_6	∞, ∞	Z_5	0
v_6	Y_6	∞	$-Z_1-Z_6$	p, 0

The integrals $v_1, \ldots, v_6$ were selected to simplify the computation of the characteristic determinant (4·6), as will be shown below; table 3 is easily derived from table 6 (Part III).

4. BOUNDARY CONDITIONS; CHARACTERISTIC EQUATION

Assume, now, that the general solution of (2·3) is

$$v = A_1 v_1 + A_2 v_2 + \ldots + A_6 v_6. \tag{4·1}$$

The boundary conditions are (2·5)

$$v = 0, \quad \frac{d^2v}{dr^2} = 0, \quad \frac{d^3v}{dr^3} - \lambda^2 \frac{dv}{dr} = 0, \tag{4·2}$$

at $r = R_1$ and $r = R_2$, since it is assumed that $\lambda_1 = \lambda$.

The following notations are introduced:

(i) $r = R_2, \quad x = x_2, \quad x > 0.$

$$\left.\begin{aligned} &\epsilon_1^{(2)} = \epsilon_4^{(2)} = \lambda^2 a^2, \quad \epsilon_2^{(2)} = \epsilon_5^{(2)} = \lambda^2 \bar{a}^2, \quad \epsilon_3^{(2)} = \epsilon_6^{(2)} = \lambda^2 a_0^2,\\ &\eta_1^{(2)} = -\eta_4^{(2)} = \lambda^3 a(a^2-1), \quad \eta_2^{(2)} = -\eta_5^{(2)} = -\lambda^3 \bar{a}(\bar{a}^2-1),\\ &\eta_3^{(2)} = -\eta_6^{(2)} = -\lambda^3 a_0(a_0^2-1). \end{aligned}\right\} \tag{4·3}$$

(ii) $r = R_1, \quad x = x_1, \quad x < 0.$

$$\left.\begin{aligned} &\epsilon_1^{(1)} = \epsilon_4^{(1)} = \lambda^2 b_0^2, \quad \epsilon_2^{(1)} = \epsilon_5^{(1)} = \lambda^2 b^2, \quad \epsilon_3^{(1)} = \epsilon_6^{(1)} = \lambda^2 \bar{b}^2,\\ &\eta_1^{(1)} = -\eta_4^{(1)} = \lambda^3 b_0(b_0^2-1), \quad \eta_2^{(1)} = -\eta_5^{(1)} = \lambda^3 b(b^2-1),\\ &\eta_3^{(1)} = -\eta_6^{(1)} = \lambda^3 \bar{b}(\bar{b}^2-1). \end{aligned}\right\} \tag{4·4}$$

It is clear that the boundary expressions (4·2) for, say, Y_1 (the outer cylinder) are

$$Y_1, \quad \epsilon_1^{(2)} Y_1, \quad \eta_1^{(2)} Y_1, \tag{4·5}$$

respectively, with the corresponding expressions for Z_1, where in the differentiations only the dominant terms were retained.

Making use of (4·1), (4·2), (4·3) and (4·4), and eliminating the constants $A_1, A_2, \ldots, A_6$, one finds the characteristic equation

$$\Delta = \begin{vmatrix} Y_1, & Y_2, & Y_2+Y_3, & \\ \epsilon_1^{(2)}Y_1, & \epsilon_2^{(2)}Y_2, & \epsilon_2^{(2)}Y_2+\epsilon_3^{(2)}Y_3, & \\ \eta_1^{(2)}Y_1, & \eta_2^{(2)}Y_2, & \eta_2^{(2)}Y_2+\eta_3^{(2)}Y_3, & \\ -Z_1, & -Z_2+Z_1, & Z_3+Z_1, & \\ -\epsilon_1^{(1)}Z_1, & -\epsilon_2^{(1)}Z_2+\epsilon_1^{(1)}Z_1, & \epsilon_3^{(1)}Z_3+\epsilon_1^{(1)}Z_1, & \\ -\eta_1^{(1)}Z_1, & -\eta_2^{(1)}Z_2+\eta_1^{(1)}Z_1, & \eta_3^{(1)}Z_3+\eta_1^{(1)}Z_1, & \\ & Y_4, & Y_5-Y_6, & Y_6 \\ & \epsilon_4^{(2)}Y_4, & \epsilon_5^{(2)}Y_5-\epsilon_6^{(2)}Y_6, & \epsilon_6^{(2)}Y_6 \\ & \eta_4^{(2)}Y_4, & \eta_5^{(2)}Y_5-\eta_6^{(2)}Y_6, & \eta_6^{(2)}Y_6 \\ & Z_4-Z_3, & Z_5, & -Z_1-Z_6 \\ & \epsilon_4^{(1)}Z_4-\epsilon_3^{(1)}Z_3, & \epsilon_5^{(1)}Z_5, & -\epsilon_1^{(1)}Z_1-\epsilon_6^{(1)}Z_6 \\ & \eta_4^{(1)}Z_4-\eta_3^{(1)}Z_3, & \eta_5^{(1)}Z_5, & -\eta_1^{(1)}Z_1-\eta_6^{(1)}Z_6 \end{vmatrix} = 0. \tag{4·6}$$

As can be seen from table 3 the integrals $v_1 \dots v_6$ are so selected that a large Y corresponds to a small, or periodic Z, and vice versa; accordingly, if only terms of the highest order are retained, the determinant (4·6) can be transformed, by suitably rearranging the columns, into products of two factors, namely,

$$\Delta = \begin{vmatrix} Y_1, & Y_5 - Y_6, & Y_6 \\ \epsilon_1^{(2)} Y_1, & \epsilon_5^{(2)} Y_5 - \epsilon_6^{(2)} Y_6, & \epsilon_6^{(2)} Y_6 \\ \eta_1^{(2)} Y_1, & \eta_5^{(2)} Y_5 - \eta_6^{(2)} Y_6, & \eta_6^{(2)} Y_6 \end{vmatrix}$$

$$\begin{vmatrix} -Z_2 + Z_1, & Z_3 + Z_1, & Z_4 - Z_3 \\ -\epsilon_2^{(1)} Z_2 + \epsilon_1^{(1)} Z_1, & \epsilon_3^{(1)} Z_3 + \epsilon_1^{(1)} Z_1, & \epsilon_4^{(1)} Z_4 - \epsilon_3^{(1)} Z_3 \\ -\eta_2^{(1)} Z_2 + \eta_1^{(1)} Z_1, & \eta_3^{(1)} Z_3 + \eta_1^{(1)} Z_1, & \eta_4^{(1)} Z_4 - \eta_3^{(1)} Z_3 \end{vmatrix} = 0, \qquad (4·7)$$

and, since the first determinant is not equal to zero, the second determinant has to vanish, whence

$$\begin{vmatrix} -Z_2 + Z_1, & Z_3 + Z_1, & Z_4 - Z_3 \\ -\epsilon_2^{(1)} Z_2 + \epsilon_1^{(1)} Z_1, & \epsilon_3^{(1)} Z_3 + \epsilon_1^{(1)} Z_1, & \epsilon_4^{(1)} Z_4 - \epsilon_3^{(1)} Z_3 \\ -\eta_2^{(1)} Z_2 + \eta_1^{(1)} Z_1, & \eta_3^{(1)} Z_3 + \eta_1^{(1)} Z_1, & \eta_4^{(1)} Z_4 - \eta_3^{(1)} Z_3 \end{vmatrix} = 0. \qquad (4·8)$$

Adding the second column to the third one, neglecting Z_1 in comparison with Z_2 and Z_3 in the remaining columns, then, after few simplifications,

$$Z_2 Z_3 \begin{vmatrix} 1, & 1, & Z_4 + Z_1 \\ \epsilon_2^{(1)}, & \epsilon_3^{(1)}, & \epsilon_1^{(1)}(Z_4 + Z_1) \\ \eta_2^{(1)}, & \eta_3^{(1)}, & -\eta_1^{(1)}(Z_4 - Z_1) \end{vmatrix} = 0, \qquad (4·9)$$

whence, evaluating the determinant,

$$\left\{ \frac{\eta_3^{(1)}}{\eta_1^{(1)}} \frac{\epsilon_2^{(1)} - \epsilon_1^{(1)}}{\epsilon_3^{(1)} - \epsilon_2^{(1)}} - \frac{\eta_2^{(1)}}{\eta_1^{(1)}} \frac{\epsilon_3^{(1)} - \epsilon_1^{(1)}}{\epsilon_3^{(1)} - \epsilon_2^{(1)}} \right\} (Z_4 + Z_1) = Z_4 - Z_1. \qquad (4·10)$$

Making use of table 2 and (4·4), then

$$\frac{\epsilon_2^{(1)} - \epsilon_1^{(1)}}{\epsilon_3^{(1)} - \epsilon_2^{(1)}} = -e^{\frac{1}{3}\pi i}, \quad \frac{\epsilon_3^{(1)} - \epsilon_1^{(1)}}{\epsilon_3^{(1)} - \epsilon_2^{(1)}} = e^{-\frac{1}{3}\pi i}, \quad \frac{\eta_3^{(1)}}{\eta_1^{(1)}} = \frac{\bar{b}}{b_0} e^{-\frac{2}{3}\pi i}, \quad \frac{\eta_2^{(1)}}{\eta_1^{(1)}} = \frac{b}{b_0} e^{\frac{2}{3}\pi i}. \qquad (4·11)$$

Now (table 2)

$$b_0 = [1 + e^{\pi i}(-hx_1)^{\frac{1}{3}}]^{\frac{1}{2}} = e^{\frac{1}{2}\pi i}[(-hx_1)^{\frac{1}{3}} - 1]^{\frac{1}{2}} = e^{\frac{1}{2}\pi i}\, l, \qquad (4·12)$$

where the letter l is introduced for brevity; and let

$$b = [1 + e^{\frac{5}{3}\pi i}(-hx_1)^{\frac{1}{3}}]^{\frac{1}{2}} = K e^{ik}, \qquad (4·13)$$

where $k \to \frac{5}{6}\pi$ if $|x_1| \to \infty$, whence

$$\bar{b} = -[1 + e^{\frac{1}{3}\pi i}(-hx_1)^{\frac{1}{3}}]^{\frac{1}{2}} = K e^{-ik}, \quad K = [1 + (-hx_1)^{\frac{1}{3}} + (-hx_1)^{\frac{2}{3}}]^{\frac{1}{4}} \qquad (4·14)$$

and finally

$$\frac{\eta_3^{(1)}}{\eta_1^{(1)}} \frac{\epsilon_2^{(1)} - \epsilon_1^{(1)}}{\epsilon_3^{(1)} - \epsilon_2^{(1)}} - \frac{\eta_2^{(1)}}{\eta_1^{(1)}} \frac{\epsilon_3^{(1)} - \epsilon_1^{(1)}}{\epsilon_3^{(1)} - \epsilon_2^{(1)}} = -\frac{2}{il} K \cos(k + \tfrac{1}{3}\pi). \qquad (4·15)$$

Now evaluate the ratio
$$\frac{Z_4-Z_1}{Z_4+Z_1}. \tag{4·16}$$

Hence from (2·21) and table 2
$$\xi_1 = i\lambda L, \quad \xi_4 = -i\lambda L, \quad \xi_4' = \xi_1' e^{\pi i}, \tag{4·17}$$
whence (table 2)
$$Z_1 \sim -e^{i\lambda L}, \quad Z_4 \sim -e^{-\frac{1}{2}\pi i}\, e^{-i\lambda L}. \tag{4·18}$$
Substituting in (4·16), then
$$\frac{Z_4-Z_1}{Z_4+Z_1} = -i\tan(\lambda L+\tfrac{1}{4}\pi). \tag{4·19}$$
From (4·10), (4·15) and (4·19) the characteristic equation follows:
$$\tan(\lambda L+\tfrac{1}{4}\pi) = -\frac{2K}{l}\cos(k+\tfrac{1}{3}\pi). \tag{4·20}$$

5. Solution of the characteristic equation

Collecting the results obtained, the question is reduced to the solution of the following two equations:
$$\tan(\lambda L+\tfrac{1}{4}\pi) = -\frac{2K}{l}\cos(k+\tfrac{1}{3}\pi), \quad h = \frac{8\Omega_1^2}{\nu^2\lambda^4}\frac{(1-\mu R_2^2/R_1^2)^2}{(R_2^2/R_1^2-1)^2 R_0}, \tag{5·1}$$
where
$$\left.\begin{aligned} R_0 &= \frac{(1-\mu)^{\frac{1}{2}} R_2}{(1-\mu R_2^2/R_1^2)^{\frac{1}{2}}},\\ L &= -\frac{6}{h}[(-hx_1)^{\frac{1}{3}}-1]^{\frac{3}{2}}\left\{\frac{(-hx_1)^{\frac{2}{3}}}{7}+\tfrac{4}{35}(-hx_1)^{\frac{1}{3}}+\tfrac{8}{105}\right\},\\ Ke^{ik} &= [1+e^{\frac{2}{3}\pi i}(-hx_1)^{\frac{1}{3}}]^{\frac{1}{2}}, \end{aligned}\right\} \tag{5·2}$$
where $k \to \frac{5}{6}\pi$ when $|hx| \to \infty$
$$l = [(-hx_1)^{\frac{1}{3}}-1]^{\frac{1}{2}} > 0,$$
and (4·14)
$$K = [1+(-hx_1)^{\frac{1}{3}}+(-hx_1)^{\frac{2}{3}}]^{\frac{1}{4}}. \tag{5·3}$$
R_0 is the radius where the mean velocity V vanishes, and x_1 is the value of x on the inner cylinder, i.e.
$$x_1 = R_1 - R_0. \tag{5·4}$$

The two equations (5·1) in the three unknown quantities Ω_1, h and λ have to be solved under the additional condition that Ω_1 is minimum for variable h; i.e. under the condition that
$$\frac{d\Omega_1}{dh} \sim \frac{d(h^{\frac{1}{2}}\lambda^2)}{dh} = 0. \tag{5·5}$$

As a first, and a very rough approximation, one can put
$$L \sim [(-hx_1)^{\frac{1}{3}}-1]^{\frac{1}{2}} x_1, \tag{5·6}$$
where terms of order $(-hx_1)^{\frac{1}{3}}$ were neglected; it is also assumed that the right-hand side of the first equation (5·1) changes slowly with the change of h; accordingly it is necessary to find the minimum of
$$\Omega_1 \sim h^{\frac{1}{2}}\lambda^2 \tag{5·7}$$

for variable h, under the condition that (5·6)

$$\lambda L \sim [(-hx_1)^{\frac{1}{3}} - 1]^{\frac{1}{2}} \lambda = \text{const.}, \tag{5·8}$$

whence

$$(-hx_1)^{\frac{1}{3}} = 3. \tag{5·9}$$

To find a better approximation, assume a set of values for $(-hx_1)^{\frac{1}{3}}$, starting with the above (5·9); from the first equation (5·1) is found the corresponding λ; the value of Ω_1 is then obtained from the second equation (5·1).

Plotting Ω_1 against $(-hx)^{\frac{1}{3}}$ then the $(-hx_1)^{\frac{1}{3}}$ corresponding to minimum of Ω_1 can be found.

Once the required value of $(-hx_1)^{\frac{1}{3}}$ is obtained, the general expression for λ and Ω_1 follows; whence their values for different radii and different μ can be evaluated.

In order that the equation (5·1) may have a real solution L must be real, i.e. the condition $|hx_1| > 1$ has to be satisfied; that puts an upper limit on λ.

6. NUMERICAL RESULTS. COMPARISON WITH EXPERIMENT

Below are given the results of these computations; the value of $(-hx_1)^{\frac{1}{3}}$ corresponding to minimum of Ω_1 is

$$(-hx_1)^{\frac{1}{3}} = 4·05. \tag{6·1}$$

Denoting by d_1 the distance between the inner cylinder and the corresponding point where the mean velocity vanishes, i.e.

$$d_1 = -x_1 = R_0 - R_1, \tag{6·2}$$

then

$$\lambda d_1 = \frac{\pi}{1·546}, \quad h = \frac{4·05^3}{d_1} = \frac{66·43}{d_1} = \frac{8\Omega_1^2}{\nu^2\lambda^4}\frac{(1-\mu R_2^2/R_1^2)^2}{(R_2^2/R_1^2-1)^2 R_0}. \tag{6·3}$$

The corresponding relations in the case when the cylinders rotate in the same direction are

$$\lambda d = \pi, \quad \frac{4\Omega_1^2}{\nu^2\lambda^4}\frac{(1-\mu R_2^2/R_1^2)}{1-R_2^2/R_1^2}\,\frac{1+\mu-\frac{3}{2}\frac{r_0}{R_0}(1-\mu)}{2} = -17·576,$$

where d is the distance between the cylinders, R_0 is now the mean radius and $2r_0 = R_2 - R_1$.

The computation leads also to the result that

$$|\lambda L| \eqsim 2·8. \tag{6·4}$$

Now, it has been assumed in evaluating the determinant (4·6) that certain of the Y functions are large, i.e. that the value of $|\lambda L|$ corresponding to the outer cylinder is sufficiently large.

Since L is approximately proportional to the distance of the corresponding cylinder from the point where the velocity V vanishes, the distance $R_2 - R_0$ should only be slightly less than d_1; hence the conclusion follows that *the solution is valid provided that $-\mu$ is not much smaller than half.*

The spacings of the vortices are very sensitive to fluctuations of the minimum Ω_1; for instance, when ($R_1 = 3{\cdot}80$ cm.)

$$\Omega_1/\nu = 336 \pm 2,$$

the corresponding spacings are about

$$0{\cdot}173 \pm 0{\cdot}016 \text{ cm.}$$

In fact, Taylor (1923) and Lewis (1927–8) found considerable fluctuations in the observed values of the spacings.

TABLE 4. $R_1 = 3{\cdot}00$ CM., $R_2 = 4{\cdot}035$ CM.

$-\mu$	d_1	π/λ	Ω_1/ν calculated	Ω_1/ν observed
0·565	0·550	0·850	40·0	37·2
0·60	0·543	0·826	41·6	41·4
0·703	0·493	0·763	46·3	42·3
0·905	0·429	0·664	56·0	52·0
1·073	0·388	0·599	64·4	62·8
1·285	0·345	0·534	75·6	79·6
1·48	0·314	0·485	86·5	87·9
1·67	0·288	0·445	97·5	97·0
2·00	0·252	0·390	117·7	118·0
2·37	0·221	0·342	141·9	137·0

TABLE 5. $R_1 = 3{\cdot}55$ CM., $R_2 = 4{\cdot}035$ CM.

$-\mu$	d_1	π/λ	Ω_1/ν calculated	Ω_1/ν observed	Ω_1/ν calculated by Taylor
0·479	0·307	0·474	80·8	80·7	79·0
0·585	0·284	0·439	90·4	89·9	84·8
0·591	0·283	0·437	90·8	91·0	84·0
0·689	0·264	0·409	98·7	96·2	—
0·793	0·247	0·382	109·8	106·5	—
0·800	0·246	0·381	110·5	105·6	—
0·843	0·240	0·371	114·8	109·4	—
1·000	0·219	0·339	130·7	128·9	—
1·129	0·205	0·317	144·0	135·9	—
1·244	0·193	0·299	156·5	148·0	—
1·302	0·188	0·291	163·0	161·1	—
1·347	0·184	0·285	167·9	—	172·8
1·489	0·173	0·267	184·2	177·3	—
1·63	0·163	0·252	200·7	183·7	—
1·795	0·153	0·236	220·9	209·4	—
1·925	0·146	0·225	236·6	215·0	—
2·00	0·142	0·219	246·3	237·3	—
2·17	0·134	0·207	268·2	235·9	—
2·32	0·127	0·197	288·2	249·8	—
2·53	0·119	0·184	317·2	280·2	—
2·68	0·114	0·177	337·4	306·0	—
2·84	0·109	0·169	360·2	318·0	—
3·25	0·098	0·152	421·0	393·0	—

TABLE 6. $R_1 = 3{\cdot}80$ CM., $R_2 = 4{\cdot}035$ CM.

$-\mu$	d_1	π/λ	Ω_1/ν calculated	Ω_1/ν observed	Ω_1/ν calculated by Taylor
0·46	0·156	0·241	206·7	219·0	209·0
0·553	0·147	0·227	224·0	219·1	222·0
0·621	0·140	0·216	241·6	227·0	230·0
1·00	0·112	0·173	336·1	320·0	—
1·16	0·103	0·159	381·7	345·3	—
1·26	0·099	0·153	401·7	367·0	—
1·36	0·094	0·146	431·9	396·6	—
1·428	0·092	0·142	450·1	415·5	—
1·500	0·089	0·138	470·5	—	475·0
1·605	0·085	0·132	502·7	447·5	—
1·996	0·073$_8$	0·114	620·7	553·0	—
2·51	0·062$_7$	0·097	791·4	666·0	—
2·891	0·056$_4$	0·087	826·7	733·0	—

TABLE 7. SPACINGS OF VORTICES. $R_1 = 3{\cdot}80$ CM., $R_2 = 4{\cdot}035$ CM.

$-\mu$	calculated	observed by Taylor
0·640	0·213	0·232, 0·228
0·716	0·204	0·201, 0·203, 0·198
1·000	0·173	0·150, 0·165, 0·160, 0·157
1·20	0·158	0·156, 0·165
1·37	0·145	0·143, 0·146
1·78	0·124	0·09, 0·105, 0·115

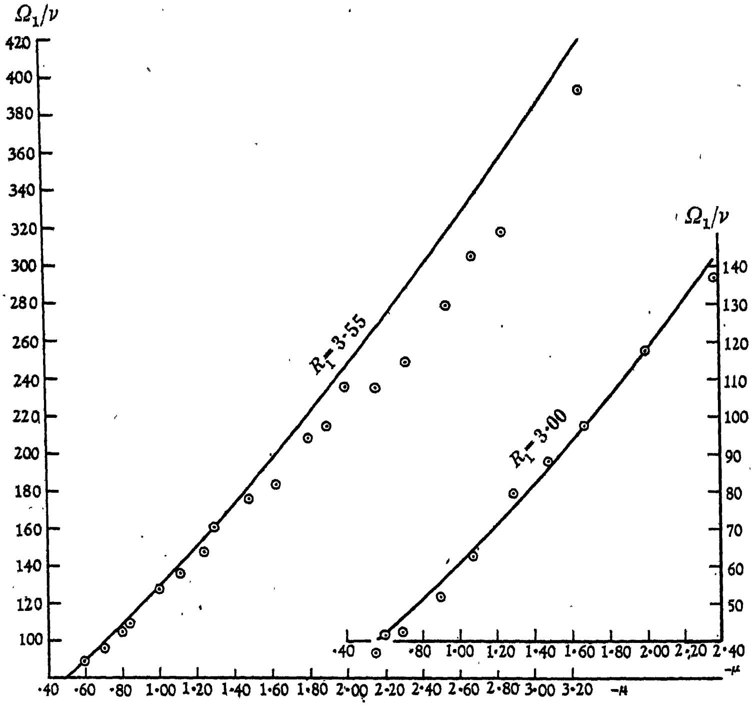


FIGURE 1. ⊙ = Observed points.

In tables 4–7 and figures 1 and 2 are given the calculated and Taylor's observed values of the critical Ω_1/ν, and the spacings.

Since Ω_1 is proportional to $d_1^{-\frac{3}{2}}$, the correct value of R_1 is very important, particularly in the case of small (R_2-R_1) and large $-\mu$.

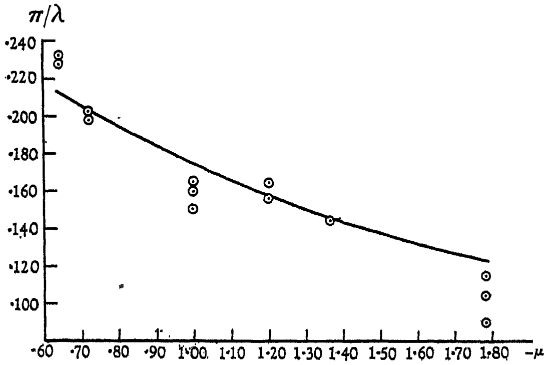


FIGURE 2. Spacings of vortices (π/λ); $R_1 = 3{\cdot}80$ cm.
⊙ = Observed points.

For that reason the agreement between the observed and calculated values of Ω_1 becomes worse when (R_2-R_1) gets smaller, and $-\mu$ larger.

The agreement is very close for the cases when $R_1 = 3{\cdot}00$ cm. and $R_1 = 3{\cdot}55$ cm., and rather poor for the case when $R_1 = 3{\cdot}80$ cm. and large values of $-\mu$.

The measured value $R_1 = 3{\cdot}80$ cm. is, it seems, not accurate enough; a small change in the latter, namely, $R_1 = 3{\cdot}79$ cm., would account for the greater part of the difference between the observed and calculated values of Ω_1. Sir Geoffrey Taylor has kindly pointed out to me that the poor agreement in this case is probably due to a slight error in alinement of the outer cylinder.

As can be seen from tables 5 and 6 our critical values of Ω_1, namely,

$$\Omega_1/\nu = 167{\cdot}9, \quad \Omega_1/\nu = 470{\cdot}5,$$

are in close agreement with the corresponding calculated results of Taylor; so are the values of the spacings, as can be seen from table 8, where Taylor's values for the spacings were found from the equation (Taylor 1923, p. 321)

$$\lambda d = 1{\cdot}73\pi,$$

d being the distance between the cylinders.

TABLE 8

R_1	$-\mu$	spacings (6·3)	Taylor's value
3·80	1·500	0·138	0·136
3·55	1·347	0·285	0·280

REFERENCES

Lewis, J. W. 1927–8 *Proc. Roy. Soc.* A, **117**, 388.
Meksyn, D. 1946 *Proc. Roy. Soc.* A, **187**, 115.
Taylor, G. I. 1923 *Phil. Trans.* A, **223**, 289.

Stability of viscous flow between rotating cylinders. III. Integration of a sixth order linear equation

By D. Meksyn, D.Sc.

(*Communicated by Professor G. Temple, F.R.S.—Received* 28 *February* 1946)

1. Introduction

The aim of this paper is to derive the asymptotic integrals, and their transformations through the critical points, of a certain linear differential equation of the sixth order containing a large parameter.

This particular equation is of importance in connexion with the question of stability of viscous flow between rotating cylinders.

Since, however, similar equations occur in all questions of stability of viscous flow, a development of proper methods of solution of such equations is of very great importance for problems of viscous flow at high Reynolds numbers.

The method of finding asymptotic integrals of linear differential equations containing a large parameter is well known; it was developed by Horn (1899), Schlesinger (1907), Birkhoff (1908) and Fowler & Lock (1922).

The main difficulty of the problem consists in the following.

The coefficients of the differential equation are expressions like $\lambda\phi(x)$, where λ is a large parameter, and $\phi(x)$ is a slowly varying function of the independent variable; the function $\phi(x)$ usually vanishes within the range of x under consideration, with the result that the asymptotic expansions become infinite at such critical points, lose their validity round these points and change their form in passing through such points.

The main problem of integration consists, thus, in finding the transformations of the asymptotic integrals in passing through critical points.

This problem was considered by Jeffreys (1924, 1942), Kramers (1926) and Goldstein (1928, 1932) for certain second-order equations.

Langer (1931), using a different method, discussed several cases of second-order equations; a summary of methods used and results obtained was also given by Langer (1934).

A case of a fourth-order equation was solved by Meksyn (in Press).

2. Method of solution

In order to explain the idea of the method, consider the case treated by Jeffreys (1924).

Integrate
$$y'' + \lambda\phi(x)\,y = 0, \tag{2·1}$$

where λ is a large parameter, and $\phi(x)$ is a slowly varying, bounded function of x.

To solve (2·1), assume that
$$y = Ae^{\xi(x)}, \tag{2·2}$$
where A is a slowly varying function of x, and ξ contains the large parameter.

It is easily found that the asymptotic solution of (2·1) is
$$y = [\phi(x)]^{-\frac{1}{4}} \exp\left\{\pm i\lambda^{\frac{1}{2}} \int^{x} [\phi(x)]^{\frac{1}{2}}\, dx\right\}. \tag{2·3}$$

Let $\phi(x)$ vanish at $x = 0$; it is clear that y becomes infinite at that point, which is, accordingly, a critical point of the expansion; and in passing through this critical point, the integral becomes transformed into a linear combination, with constant coefficients, of two particular integrals.

The problem consists in finding these coefficients.

To that end, expand $\phi(x)$ round the critical point, and let
$$\phi(x) = \phi'(0)\,x + \tfrac{1}{2}\phi''(0)\,x^2 + \ldots, \tag{2·4}$$
where $\phi'(0) \neq 0$.

Consider the auxiliary equation
$$y'' + \lambda\phi'(0)\,xy = 0. \tag{2·5}$$

If
$$\frac{1}{2}\frac{\phi''(0)\,x}{\phi'(0)} \ll 1, \tag{2·6}$$
the two equations (2·1) and (2·5) will behave similarly round the critical point $x = 0$; and if the law of transformation can be found for the integrals of (2·5), the corresponding law of transformation can be derived for equation (2·1). In the present case, (2·5) can be solved in Bessel functions.

It is clear, therefore, that for the success of this method it is necessary that λ should be sufficiently large, so that λx may be large for sufficiently small values of x; and that $\phi(x)$ should be a slowly varying function with a limited variation.

The problem, therefore, consists in the following. It is necessary:

A. To find the asymptotic integrals of a certain equation, and their critical points.

B. To derive the auxiliary equation, integrate it, and find the transformations of its asymptotic integrals.

C. To find the transformations of the asymptotic integrals of the original equation, and to derive the expressions for the integrals close to the critical points.

A. ASYMPTOTIC INTEGRALS OF THE ORIGINAL EQUATION

3. Asymptotic integrals

Integrate the differential equation (Part II, (2·3))
$$\begin{aligned}
&\frac{d^6\psi}{dr^6} + \frac{3}{r}\frac{d^5\psi}{dr^5} - \left(\frac{6}{r^2} + 2\lambda_1^2 + \lambda^2\right)\frac{d^4\psi}{dr^4} + \left(\frac{12}{r^3} - \frac{4\lambda_1^2 + 2\lambda^2}{r}\right)\frac{d^3\psi}{dr^3} \\
&\quad + \left(-\frac{27}{r^4} + \frac{6\lambda_1^2 + 3\lambda^2}{r^2} + \lambda_1^4 + 2\lambda^2\lambda_1^2\right)\frac{d^2\psi}{dr^2} + \left(\frac{45}{r^5} - \frac{6\lambda_1^2 + 3\lambda^2}{r^3} + \frac{\lambda_1^4 + 2\lambda^2\lambda_1^2}{r}\right)\frac{d\psi}{dr} \\
&\quad + \left[-\frac{45}{r^6} + \frac{6\lambda_1^2 + 3\lambda^2}{r^4} - \frac{\lambda_1^4 + 2\lambda^2\lambda_1^2}{r^2} - \lambda^2\lambda_1^4 - \frac{4\lambda^2}{r^2}\left(A^2 + \frac{AB}{r^2}\right)\right]\psi = 0,
\end{aligned} \tag{3·1}$$

where λ and λ_1 are two large parameters, which are of the same order of magnitude, and A^2/ν^2 and AB/ν^2 are constants of the order of magnitude λ^4; r is the independent variable; it is positive, of order of unity, and it changes round a value $r = R_0$, so that

$$R_0 - \varDelta R_0 \leqslant r < R_0 + \varDelta R_0, \quad \varDelta R_0/R_0 \ll 1,$$

i.e. r does not reach the zero point.

To integrate (3·1) assume
$$\psi = Ce^{\xi}, \tag{3·2}$$
where ξ is a function of r, and it contains the large parameter, and C is a slowly varying function of r.

Substituting (3·2) in (3·1) and equating to zero terms of order λ^6 and λ^5 respectively, then

$$Cf(\xi') = 0, \quad C'\frac{df(\xi')}{d\xi'} + \frac{1}{2}\frac{C}{r}\frac{df(\xi')}{d\xi'} + \frac{C}{2}\frac{d^2f(\xi')}{d\xi'^2}\xi' = 0, \tag{3·3}$$

where $f(\xi') = \xi'^6 - (2\lambda_1^2 + \lambda^2)\,\xi'^4 + (\lambda_1^4 + 2\lambda^2\lambda_1^2)\,\xi'^2 - \lambda^2\lambda_1^4 - \frac{4\lambda^2}{\nu^2}\left(A^2 + \frac{AB}{r^2}\right),$

where primes denote differentiation with respect to r.

Integrating (3·3) one finds C, and the particular integrals of (3·1) are of the form

$$\psi \sim r^{-\frac{1}{2}}\left[\frac{\partial f(\xi')}{\partial \xi'}\right]^{-\frac{1}{2}} e^{\xi}, \tag{3·4}$$

where ξ' is found from the equation $f(\xi') = 0$, and ξ' is integrated between appropriate limits; there are six particular independent integrals.

In what follows it is assumed that $\lambda_1 = \lambda$, which corresponds to steady motion, in the hydrodynamical problem.

4. The critical points

Since r does not vanish within the range under consideration, the critical points are found from the condition (3·4)

$$\frac{df(\xi')}{d\xi'} = 0, \tag{4·1}$$

whence from (3·3) and (4·1)
$$\xi'(\xi'^2 - \lambda^2) = 0, \tag{4·2}$$
or, making use of the expression for $f(\xi')$ (3·3), the critical values of r are found from

$$\lambda^6 + \frac{4\lambda^2}{\nu^2}\left(A^2 + \frac{AB}{r^2}\right) = 0, \tag{4·3}$$

and
$$A^2 + \frac{AB}{r^2} = 0. \tag{4·4}$$

In the hydrodynamical problem, where this equation is important, these critical points turn out to be very close together; accordingly, the critical point is selected as that corresponding to $\xi' = 0$, i.e.

$$1 + \frac{4}{\nu^2\lambda^4}\left(A^2 + \frac{AB}{r^2}\right) = 0. \tag{4·5}$$

Let $r = R_0$ be the root of
$$A^2+\frac{AB}{r^2} = 0, \tag{4·6}$$
and put
$$r = R_0+x. \tag{4·7}$$
Retaining only terms linear in x, i.e. neglecting terms of order $(x/R_0)^2$ and higher the expression (4·5) can be written as
$$1+\frac{4}{\nu^2\lambda^4}\left(A^2+\frac{AB}{r^2}\right) = 1+hx, \tag{4·8}$$
where h is a constant.

Substituting (4·8) in $f(\xi')$ (3·3), and, solving with respect to ξ', then
$$\xi' = \pm\lambda[1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}}, \tag{4·9}$$
whence
$$\xi = \pm\lambda\int_{x_0}^{x}[1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}}\,dx, \tag{4·10}$$
where x_0 is the root of the equation
$$1+hx = 0, \tag{4·11}$$
and ω is any cube root of unity; x_0 is taken as the lower limit of integration, since ξ' has to vanish at the critical point (4·11).

To integrate (4·10), put
$$1+\omega(hx)^{\frac{1}{3}} = q^2, \tag{4·12}$$
and it can be seen that our selection of the critical point (4·11) leads to a somewhat simpler result than that to which the critical point $x = 0$ would lead; accordingly
$$\xi = \pm\frac{6\lambda}{h}[1+\omega(hx)^{\frac{1}{3}}]^{\frac{3}{2}}\left\{\frac{\omega^2(hx)^{\frac{2}{3}}}{7}-\tfrac{4}{35}\omega(hx)^{\frac{1}{3}}+\tfrac{8}{105}\right\}. \tag{4·13}$$

Making use of the expression for $f(\xi')$ (3·3), then, finally, the six particular integrals of (3·1) are given by (3·4)
$$\psi \sim r^{-\frac{1}{2}}(\xi')^{-\frac{1}{2}}\,(hx)^{-\frac{1}{3}}\,e^{\xi}, \tag{4·14}$$
where the six values of ξ' and ξ are found from (4·9) and (4·13).

B. INTEGRATION OF THE AUXILIARY EQUATION

5. The auxiliary equation

To find the auxiliary equation, it is convenient first to transform (3·1).

Put
$$\psi = y/r^{\frac{1}{2}}. \tag{5·1}$$

Substituting (5·1) in (3·1), dropping all terms with powers of r in the denominators, except the term AB/r^2, and making use of (4·8), the auxiliary equation follows:
$$\frac{d^6y}{dx^6}-3\lambda^2\frac{d^4y}{dx^4}+3\lambda^4\frac{d^2y}{dx^2}-\lambda^6[1+hx]\,y = 0, \tag{5·2}$$
where the independent variable was transformed to x, and λ_1 is put equal to λ.

The disregarded terms do not contribute to the dominant term of the asymptotic integral; in the equation (3·1) the term hx was only valid in the first approximation, for sufficiently small values of x; in the auxiliary equation (5·2) the expression $1+hx$ is valid for all values of x.

6. INTEGRATION OF THE AUXILIARY EQUATION

The equation (5·2) is of known type, and it can be solved in definite integrals.

Let
$$y = \int e^{tx}\phi(t)\,dt. \tag{6·1}$$

Substituting in (5·2) then the equation will be satisfied if $\phi(t)$ satisfies
$$(t^6 - 3\lambda^2t^4 + 3\lambda^4t^2 - \lambda^6)\,\phi(t) + \lambda^6h\frac{d\phi}{dt} = 0, \tag{6·2}$$
whence
$$\phi = \exp\left\{-\frac{1}{h\lambda^6}\left(\frac{t^7}{7} - \frac{3\lambda^2t^5}{5} + \lambda^4t^3 - \lambda^6t\right)\right\}, \tag{6·3}$$
and the solution is
$$y = \int \exp\left\{-\frac{1}{h\lambda^6}\left(\frac{t^7}{7} - \frac{3\lambda^2t^5}{5} + \lambda^4t^3 - \lambda^6t\right) + tx\right\}dt, \tag{6·4}$$
provided that the integration is taken between such limits that the expression
$$\exp\left\{-\frac{1}{h\lambda^6}\left(\frac{t^7}{7} - \frac{3\lambda^2t^5}{5} + \lambda^4t^3 - \lambda^6t\right) + tx\right\} \tag{6·5}$$
vanishes.

Assume that the path of integration starts from infinity where (6·5) vanishes, proceeds to zero and returns to infinity along a suitable different path, where (6·5) again vanishes.

In order that (6·5) may vanish at the two extremities of the path, the real part of t^7 must be positive for infinite values of t, since h can be taken as positive; the parameter λ is assumed to be real.

Let
$$\arg t = \theta \quad \text{if} \quad |t| \to \infty, \tag{6·6}$$
in order that (6·5) might vanish at infinity, the condition
$$2\pi n - \tfrac{1}{2}\pi < 7\theta < 2\pi n + \tfrac{1}{2}\pi, \tag{6·7}$$
where n is a positive integer, must be satisfied; whence
$$\frac{2\pi n}{7} - \tfrac{1}{14}\pi < \theta < \frac{2\pi n}{7} + \tfrac{1}{14}\pi. \tag{6·8}$$

7. ASYMPTOTIC INTEGRALS

In the hydrodynamical problem $|hx|^{\frac{1}{3}}$ is larger than unity on the boundaries; in what follows it is important to assess approximately the argument of ξ; the discussion could be considerably simplified by neglecting unity in comparison with $(hx)^{\frac{1}{3}}$ in

(4·9), i.e. the integrals of (5·2) are considered for very large values of x; whence, instead of the rigorous expression for y (6·4) it can be assumed that

$$y \sim \int \exp\left\{-\frac{t^7}{7h\lambda^6}+tx\right\}dt. \tag{7·1}$$

To find the asymptotic behaviour of y the path of integration is deformed so that it passes through the stationary points of the integrand, namely, through the points which satisfy

$$\frac{dF}{dt}=\frac{d}{dt}\left(-\frac{t^7}{7h\lambda^6}+tx\right)=0, \tag{7·2}$$

along the path of the steepest descent, the new notation $F(t)$ being introduced for brevity; accordingly, the stationary points are at

$$t_0=\varpi\lambda(hx)^{\frac{1}{6}}, \tag{7·3}$$

where ϖ is here the sixth root of unity.

Let

$$t=t_0+\epsilon, \tag{7·4}$$

whence, retaining only terms of the order ϵ^2,

$$\left.\begin{aligned} F(t) &= F(t_0+\epsilon)\approx F(t_0)+\tfrac{1}{2}F''(t_0)\,\epsilon^2,\\ F(t_0) &= \tfrac{6}{7}\varpi\lambda x(hx)^{\frac{1}{6}},\\ F''(t_0) &= -\frac{6}{h\lambda}\varpi^{5}(hx)^{\frac{5}{6}}. \end{aligned}\right\} \tag{7·5}$$

Combining now (7·2) with (7·5), and, extending the integration with respect to ϵ to infinity along a suitable path, the asymptotic values of y can easily be evaluated.

As an example one case will be considered in detail, and only the final results will be given. The paths of steepest descent are indicated approximately in figures 1 and 2, in which are given the general schemes of integration for positive and negative values of x, the arguments of x being zero and π respectively.

The complex plane is divided into fourteen sections (6·8) where the integrand (7·1) tends alternatively to infinity and zero.

The path of integration starts in a section where the integrand vanishes at infinity, proceeds along any suitable path to zero and returns in a different section, where the integrand again vanishes at infinity, the path being described in the anticlockwise direction.

These sections are denoted in figures 1 and 2 by I, II, ..., VII.

To find the asymptotic values of the integrals these paths are deformed as indicated in figures 1 and 2.

Consider now the case of the integral corresponding to the region III–II, and for x positive.

The stationary point is (7·3), $\quad t_0=\varpi(hx)^{\frac{1}{6}}, \qquad$ (7·6)

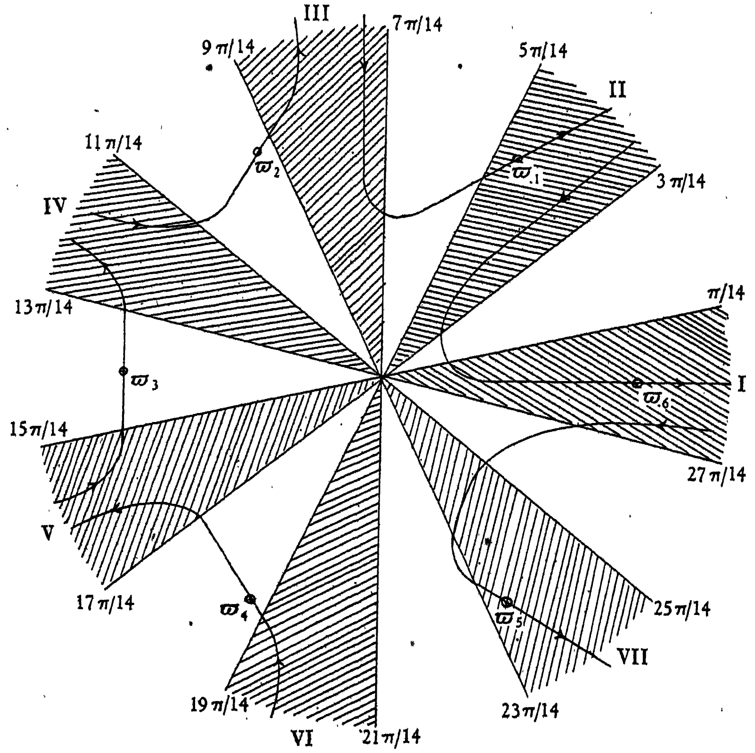


FIGURE 1. $x > 0$; arg $x = 0$.

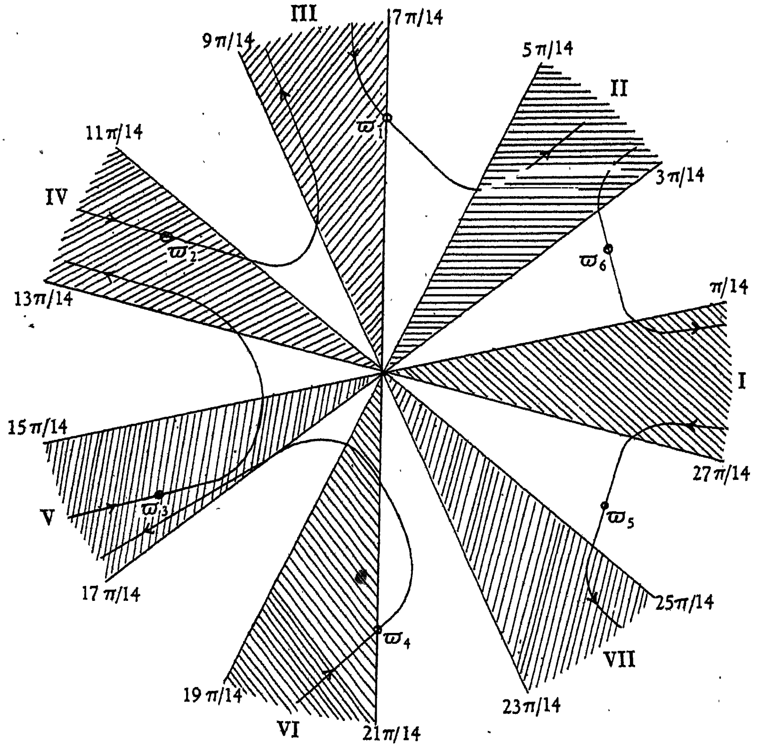


FIGURE 2. $x < 0$; arg $x = \pi$.

where ϖ is a sixth root of unity, and t_0 must lie anywhere between the sections III and II inclusive, i.e.

$$\tfrac{9}{14}\pi > \arg t_0 > \tfrac{3}{14}\pi. \tag{7·7}$$

It is evident that the only value of ϖ which satisfies these conditions is

$$\varpi_1 = e^{\frac{1}{3}\pi i}, \quad \text{whence} \quad t_0 = e^{\frac{1}{3}\pi i}(hx)^{\frac{1}{6}}. \tag{7·8}$$

It is easy to see that the path of the steepest descent is given by

$$t = t_0 + \epsilon = t_0 + \eta e^{\frac{1}{6}\pi i}, \tag{7·9}$$

where η is real; whence

$$F(t_0) = \tfrac{6}{7}e^{\frac{1}{3}\pi i}\,x(hx)^{\frac{1}{6}}, \quad F''(t_0) = -\frac{6}{h}e^{\frac{5}{3}\pi i}(hx)^{\frac{5}{6}}, \tag{7·10}$$

and the integral (7·1), which is denoted by y_1, tends to

$$y_1 = e^{\frac{6}{7}e^{\frac{1}{3}\pi i}x(hx)^{\frac{1}{6}}\lambda}\int_{-\infty}^{\infty} e^{\frac{1}{6}\pi i}\exp\left\{-\frac{3}{h\lambda}e^{2\pi i}(hx)^{\frac{5}{6}}\eta^2\right\}d\eta$$

$$= -\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{1}{6}\pi i}(hx)^{-\frac{5}{12}}\exp\left\{\tfrac{6}{7}e^{\frac{1}{3}\pi i}\lambda x(hx)^{\frac{1}{6}}\right\}. \tag{7·11}$$

In tables 1 and 2 are given the results of these computations; the integrals for negative x are denoted, for shortness, by the letter z.

Inserting in the y integrals a negative x, namely,

$$x = (-x)\,e^{\pi i}, \tag{7·12}$$

the corresponding z integrals are obtained.

The signs of z_1, z_2 and z_6 were changed in order that the corresponding y integrals may become identical with z when the substitution (7·12) is made; it should be borne in mind, however, that z_n is, generally, not a continuation of y_n through the critical point.

Table 1. $x > 0$, $\arg x = 0$, $\varpi^6 = 1$, $t_0 = \varpi\lambda(hx)^{\frac{1}{6}}$.

section	ϖ_n	y_n
II–I	$\varpi_6 = 1$	$y_6 = \sqrt{\left(\frac{\pi h\lambda}{3}\right)}(hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}\lambda x(hx)^{\frac{1}{6}}\}$
III–II	$\varpi_1 = e^{\frac{1}{3}\pi i}$	$y_1 = -\sqrt{\left(\frac{\pi h\lambda}{3}\right)}e^{\frac{1}{6}\pi i}(hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{1}{3}\pi i}\lambda x(hx)^{\frac{1}{6}}\}$
IV–III	$\varpi_2 = e^{\frac{2}{3}\pi i}$	$y_2 = \sqrt{\left(\frac{\pi h\lambda}{3}\right)}e^{\frac{1}{3}\pi i}(hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{2}{3}\pi i}\lambda x(hx)^{\frac{1}{6}}\}$
V–IV	$\varpi_3 = e^{\pi i}$	$y_3 = -\sqrt{\left(\frac{\pi h\lambda}{3}\right)}e^{\frac{1}{2}\pi i}(hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\pi i}\lambda x(hx)^{\frac{1}{6}}\}$
VI–V	$\varpi_4 = e^{\frac{4}{3}\pi i}$	$y_4 = \sqrt{\left(\frac{\pi\lambda h}{3}\right)}e^{\frac{2}{3}\pi i}(hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{4}{3}\pi i}\lambda x(hx)^{\frac{1}{6}}\}$
I–VII	$\varpi_5 = e^{\frac{5}{3}\pi i}$	$y_5 = \sqrt{\left(\frac{\pi h\lambda}{3}\right)}e^{\frac{5}{6}\pi i}(hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{5}{3}\pi i}\lambda x(hx)^{\frac{1}{6}}\}$

TABLE 2. $x<0$, $\arg x=\pi$, $\varpi^6=-1$, $t_0=\lambda\varpi(-hx)^{\frac{1}{6}}$.

section	ϖ_n	z_n
II–I	$\varpi_6=e^{\frac{1}{6}\pi i}$	$-z_6=\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{7}{12}\pi i}(-hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{1}{6}\pi i}\lambda x(-hx)^{\frac{1}{6}}\}$
III–II	$\varpi_1=e^{\frac{1}{2}\pi i}$	$-z_1=-\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{3}{4}\pi i}(-hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{1}{2}\pi i}\lambda x(-hx)^{\frac{1}{6}}\}$
IV–III	$\varpi_2=e^{\frac{5}{6}\pi i}$	$-z_2=\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{11}{12}\pi i}(-hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{5}{6}\pi i}\lambda x(-hx)^{\frac{1}{6}}\}$
V–IV	$\varpi_3=e^{\frac{7}{6}\pi i}$	$z_3=-\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{1}{12}\pi i}(-hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{7}{6}\pi i}\lambda x(-hx)^{\frac{1}{6}}\}$
VI–V	$\varpi_4=e^{\frac{3}{2}\pi i}$	$z_4=\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{1}{4}\pi i}(-hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{3}{2}\pi i}\lambda x(-hx)^{\frac{1}{6}}\}$
I–VII	$\varpi_5=e^{\frac{11}{6}\pi i}$	$z_5=\sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{5}{12}\pi i}(-hx)^{-\frac{5}{12}}\exp\{\frac{6}{7}e^{\frac{11}{6}\pi i}\lambda x(-hx)^{\frac{1}{6}}\}$

8. TRANSFORMATION OF INTEGRALS

It will now be shown, as an example, how an integral for positive values of x is transformed when x becomes negative, and, then, the final result will be given.

Consider the region II–I.

(a) $x>0$.

From figure 1 it is seen that the path of integration passes through two points, ϖ_1 and ϖ_6, and from table 1 is obtained the corresponding integral

$$y_6-y_1, \tag{8·1}$$

since the path is described in the anti-clockwise direction.

(b) $x<0$

From figure 2 it is seen that the path of integration passes only through the point ϖ_6, and from table 2 is obtained the corresponding integral

$$-z_6, \tag{8·2}$$

whence one concludes that the asymptotic integral y_6-y_1 is transformed into $-z_6$ when x changes from positive to negative values, the $\arg x$ being equal to π.

In table 3 are given the transformations of the corresponding integrals, and it is indicated whether they tend to infinity, zero or are periodic for large values of $|x|$.

As can be seen from table 3, there are seven integrals; they are not, however, independent; adding them up, it is found that the sums at the y's and the z's vanish asymptotically.

At first sight there appears a kind of ambiguity in the above transformations; for instance, from the first and second lines (table 3) it is found that y_6 is transformed to $-z_1$, whereas from the sixth and seventh lines it appears that y_6 is transformed into $-z_4$; the two functions z_1 and z_4 are not, however, asymptotically equal.

This difficulty is explained as follows: from figure 1 it can be seen the two y_6 functions are different for small values of x.

TABLE 3.

region	$x>0$, $\arg x = 0$		$x<0$, $\arg x = \pi$	
II–I	y_6-y_1	∞, ∞	$-z_6$	0
III–II	y_1	∞	$-z_1$	p
IV–III	y_2	0	$-z_2+z_1$	∞, p
V–IV	y_3	0	z_3+z_2	∞, ∞
VI–V	y_4	0	z_4-z_3	p, ∞
VII–VI	$-y_5$	∞	$-z_4$	p
I–VII	y_5-y_6	∞, ∞	z_5	0

C. TRANSFORMATION OF THE INTEGRALS OF THE ORIGINAL EQUATION

9. TRANSFORMATION OF THE INTEGRALS

It is necessary now to make use of the transformations (table 3) of the integrals of the auxiliary equation (5·2) in order to find the transformations of the integrals of the original equation (3·1).

To that end one has to find how the two sets of integrals are related.

The integrals of the equation (3·1) are all of the form

$$Ar^{-\frac{1}{2}}\left[\frac{\partial f(\xi')}{\partial \xi'}\right]^{-\frac{1}{2}} e^{\xi}, \tag{9·1}$$

where

$$\left.\begin{aligned} f(\xi') &= (\xi'^2-\lambda^2)^3-\lambda^6(1+hx),\\ \xi' &= \pm\lambda[1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}},\\ \xi &= \pm\lambda\int_{x_0}^{x}[1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}}\,dx,\\ \omega^3 &= 1, \end{aligned}\right\} \tag{9·1a}$$

and x_0 is the root of

$$1+hx=0,$$

A being a constant.

The integrals (9·1) can be transformed into

$$\left.\begin{aligned} &A_1r^{-\frac{1}{2}}(\xi')^{-\frac{1}{2}}(hx)^{-\frac{1}{3}}e^{\xi} \quad (x>0),\\ &A_2r^{-\frac{1}{2}}(\xi')^{-\frac{1}{2}}(-hx)^{-\frac{1}{3}}e^{\xi} \quad (x<0). \end{aligned}\right\} \tag{9·2}$$

Inserting the values of ξ' (9·1a) in the expressions (9·2), and comparing them with those in tables 1 and 2, it is noticed that they are identical for large values of $|x|$, except for the factor $r^{\frac{1}{2}}$ and the arbitrary constants A.

Now select the constants A, and the values of ω, so that the two sets of integrals should be equal for large values of x, except for the irrelevant factor $r^{\frac{1}{2}}$.

In tables 4 and 5 are given the values of ω and the arguments of ξ' and ξ, when $|x|$ tends to infinity, for the integrals of the original equation (3·1).

By comparing, for large x, an integral (9·2) of the original equation with a ξ (tables 4 and 5) equal to ξ of auxiliary equation (tables 1 and 2), and, bearing in mind the values of $\arg\xi'$ (tables 4 and 5), the constants A can be found so that the two integrals shall be equal.

The integrals of the original equation are denoted by the letters Y and Z, so that Y_n, Z_n correspond to y_n and z_n respectively.

TABLE 4. $x > 0$, $\arg x = 0$, $\xi = \pm\lambda\int_{x_0}^{x}[1+\omega(hx)^{\frac{1}{3}}]^{\frac{1}{2}}\,dx$, $\arg\xi' = \arg\xi$

Y_n	ω	$\arg\xi_n$	Y_n	ω	$\arg\xi_n$
Y_1	$e^{\frac{2}{3}\pi i}$	$\pi/3$	Y_4	$e^{\frac{8}{3}\pi i}$	$4\pi/3$
Y_2	$e^{\frac{4}{3}\pi i}$	$2\pi/3$	Y_5	$e^{\frac{10}{3}\pi i}$	$5\pi/3$
Y_3	$e^{2\pi i}$	π	Y_6	1	0

TABLE 5. $x<0$, $\arg x = \pi$, $\xi = \pm\lambda\int_{x_0}^{x}[1+\omega e^{\frac{1}{3}\pi i}(-hx)^{\frac{1}{3}}]^{\frac{1}{2}}\,dx$, $\arg\xi = \arg\xi' + \pi$.

Z_n	ω	$\arg\xi_n$	Z_n	ω	$\arg\xi_n$
Z_1	$e^{\frac{2}{3}\pi i}$	$3\pi/2$	Z_4	$e^{\frac{8}{3}\pi i}$	$5\pi/2$
Z_2	$e^{\frac{4}{3}\pi i}$	$11\pi/6$	Z_5	$e^{\frac{10}{3}\pi i}$	$17\pi/6$
Z_3	$e^{2\pi i}$	$13\pi/6$	Z_6	1	$7\pi/6$

We now show, as an example, how the constant A is found; and, then, give the final results.

Consider Y_2, where
$$Y_2 = \frac{Ae^{\xi_2}}{r^{\frac{1}{2}}(\xi_2')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}. \tag{9·3}$$

From table 4
$$\arg\xi_2' = \tfrac{2}{3}\pi, \tag{9·4}$$

whence
$$Y_2 \to \frac{Ae^{-\frac{1}{3}\pi i}\,e^{\xi_2}}{r^{\frac{1}{2}}(hx)^{\frac{5}{12}}}. \tag{9·5}$$

the corresponding integral y_2 (table 1) is
$$y_2 \to \sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{1}{3}\pi i}\,(hx)^{-\frac{5}{12}}\,e^{\xi_2}. \tag{9·6}$$

Comparing (9·5) and (9·6) then
$$A = \sqrt{\left(\frac{\pi h\lambda}{3}\right)}\,e^{\frac{2}{3}\pi i}. \tag{9·7}$$

The following are the particular integrals for positive x:

$$x>0, \quad \arg x = 0.$$

$$\left.\begin{aligned}
Y_1 &= \frac{e^{-\frac{2}{3}\pi i}\,e^{\xi_1}}{r^{\frac{1}{2}}(\xi_1')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, & Y_4 &= \frac{-\,e^{\frac{1}{3}\pi i}\,e^{\xi_4}}{r^{\frac{1}{2}}(\xi_4')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}},\\
Y_2 &= \frac{e^{\frac{2}{3}\pi i}\,e^{\xi_2}}{r^{\frac{1}{2}}(\xi_2')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, & Y_5 &= \frac{-\,e^{\frac{2}{3}\pi i}\,e^{\xi_5}}{r^{\frac{1}{2}}(\xi_5')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}},\\
Y_3 &= \frac{e^{\xi_3}}{r^{\frac{1}{2}}(\xi_3')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}, & Y_6 &= \frac{e^{\xi_6}}{r^{\frac{1}{2}}(\xi_6')^{\frac{1}{2}}\,(hx)^{\frac{1}{3}}}.
\end{aligned}\right\} \tag{9·8}$$

The particular integrals for negative x are

$$x<0, \quad \arg x = \pi$$

$$\left.\begin{aligned} -Z_1 &= \frac{e^{\xi_1}}{r^{\frac{1}{2}}(\xi_1')^{\frac{1}{2}}\,(-hx)^{\frac{1}{3}}}, & Z_4 &= \frac{-e^{\xi_4}}{r^{\frac{1}{2}}(\xi_4')^{\frac{1}{2}}\,(-hx)^{\frac{1}{3}}}, \\ -Z_2 &= \frac{-e^{\frac{1}{3}\pi i}\,e^{\xi_2}}{r^{\frac{1}{2}}(\xi_2')^{\frac{1}{2}}\,(-hx)^{\frac{1}{3}}}, & Z_5 &= \frac{-e^{\frac{1}{3}\pi i}\,e^{\xi_5}}{r^{\frac{1}{2}}(\xi_5')^{\frac{1}{2}}\,(-hx)^{\frac{1}{3}}}, \\ Z_3 &= \frac{e^{-\frac{1}{3}\pi i}\,e^{\xi_3}}{r^{\frac{1}{2}}(\xi_3')^{\frac{1}{2}}\,(-hx)^{\frac{1}{3}}}, & -Z_6 &= \frac{e^{\frac{2}{3}\pi i}\,e^{\xi_6}}{r^{\frac{1}{2}}(\xi_6')^{\frac{1}{2}}\,(-hx)^{\frac{1}{3}}}, \end{aligned}\right\} \tag{9·9}$$

where in (9·8) and (9·9) the irrelevant factor $(\pi h\lambda/3)^{\frac{1}{2}}$ was disregarded; it has, however, to be accounted for in the expansions for small values of x (10·2).

Making use of table 3 the transformations of the above integrals are immediately found.

TABLE 6. TRANSFORMATION OF THE INTEGRALS

$x>0$, $\arg x = 0$	$x<0$, $\arg x = \pi$	$x>0$, $\arg x = 0$	$x<0$, $\arg x = \pi$
$Y_6 - Y_1$	$-Z_6$	Y_4	$Z_4 - Z_3$
Y_1	$-Z_1$	Y_5	Z_4
Y_2	$-Z_2 + Z_1$	$Y_5 - Y_6$	Z_5
Y_3	$Z_3 + Z_2$		

10. EXPANSIONS VERY CLOSE TO THE CRITICAL POINT

It remains to show how the integrals are found for very small values of x, where the asymptotic expansions are no more valid.

The integrals of the original and the auxiliary equations are connected by (5·2)

$$Y = \frac{y}{r^{\frac{1}{2}}}. \tag{10·1}$$

Consider now Y_1; the integral y_1 of the auxiliary equation can be utilized; the path of integration may be selected along a straight line within region III, inclined, say, at an angle $\frac{4}{7}\pi$, from infinity to zero; and, then along any straight line within the region II back to infinity; for example, along a line inclined at an angle $\frac{2}{7}\pi$.

Accordingly, one obtains (6·4), bearing in mind the direction of the path,

$$Y_1 = \left(\frac{3}{\pi h\lambda}\right)^{\frac{1}{2}} \frac{1}{r^{\frac{1}{2}}} \left\{\int_0^{\infty e^{\frac{2}{7}\pi i}} \exp\{\phi(t)+tx\}\,dt - \int_0^{\infty e^{\frac{4}{7}\pi i}} \exp\{\phi(t)+tx\}\,dt\right\}, \tag{10·2}$$

where
$$\phi(t) = -\frac{1}{h\lambda^6}\left\{\frac{t^7}{7} - \tfrac{3}{5}\lambda^2 t^5 + \lambda^4 t^3 - \lambda^6 t\right\}. \tag{10·3}$$

Expanding now (10·2) in powers of x, then

$$Y_1 = \left(\frac{3}{\pi h\lambda}\right)^{\frac{1}{2}} \left\{\sum_0^\infty \left[\int_0^{\infty e^{\frac{2}{7}\pi i}} e^{\phi(t)} t^n\,dt - \int_0^{\infty e^{\frac{4}{7}\quad i}} e^{\phi(t)} t^n\,dt\right] \frac{x^n}{n\,!}\right\}, \tag{10·4}$$

where the factor $(3/\pi h\lambda)^{\frac{1}{2}}$ was introduced to compensate its disregarding in the asymptotic integrals (9·8) and (9·9).

I wish to thank Professor G. Temple, F.R.S., for his interest and encouragement, and the Department of Scientific and Industrial Research for a grant which enabled me to complete this work.

REFERENCES

Birkhoff, G. D. 1908 *Trans. Amer. Math. Soc.* **9**, 219.
Fowler, R. H. & Lock, C. N. H. 1922 *Proc. Lond. Math. Soc.* (2), **20**, 127.
Goldstein, S. 1928 *Proc. Lond. Math. Soc.* (2), **28**, 81.
Goldstein, S. 1932 *Proc. Lond. Math. Soc.* (2), **33**, 246.
Horn, J. 1899 *Math. Ann.* (52), **271**, 340.
Jeffreys, H. 1924 *Proc. Lond. Math. Soc.* **23**, 428.
Jeffreys, H. 1942 *Phil. Mag.* **33**, 451.
Kramers, H. A. 1926 *Z. Phys.* **39**, 828.
Langer, R. E. 1931 *Trans. Amer. Math. Soc.* **33**, 23 (and subsequent papers).
Langer, R. E. 1934 *Bull. Amer. Math. Soc.* **40**, 545.
Meksyn, D. *Proc. Lond. Math. Soc.* (in the Press).
Schlesinger, L. 1907 *Math. Ann.* **63**, 277.

INDEX TO VOLUME 187 (A)

Adsorption of vapours on mercury. I. Non-polar substances (Kemball & Rideal), 53.
Adsorption of vapours on mercury. II. The entropy and heat of adsorption of non-polar substances (Kemball), 73.
Algebra related to elementary particles of spin $\frac{3}{2}$ (Madhavarao, Thiruvenkatachar & Venkatachaliengar), 385.
Andrade, E. N. da C. & Dodd, C. The effect of an electric field on the viscosity of liquids, 296.
Application of quantum electrodynamics to multiple processes (Jayaratnam Eliezer), 210.

Bagnold, R. A. Motion of waves in shallow water. Interaction between waves and sand bottoms, 1.
Basic reactions in the upper atmosphere, I (Bates & Massey), 261.
Bates, D. R. & Massey, H. S. W. The basic reactions in the upper atmosphere, I, 261.
Bhagavantam, S. & Bhimasenachar, J. Elastic constants of diamond, 381.
Bhimasenachar, J. *See* Bhagavantam & Bhimasenachar.
Bibhutibhusan Sen. Two-dimensional boundary-value problems of elasticity, 87

Chalmers, B., Forrester, P. G. & Phelps, E. F. Kinetic friction in or near the boundary region. I. Apparatus and experimental methods, 430.
Cloud-chamber investigation of penetrating showers (Rochester), 464.

Dirac's theory of quantum electrodynamics: the interaction of an electron and a radiation field (Jayaratnam Eliezer), 197.
Distribution of stress in the neighbourhood of a crack in an elastic solid (Sneddon), 229.
Dodd, C. *See* Andrade & Dodd.

Effect of an electric field on the viscosity of liquids (Andrade & Dodd), 296
Elastic constants of diamond (Bhagavantam & Bhimasenachar), 381.
Experimental study of the blast from bombs and bare charges (Grime & Sheard), 357.

Forrester, P. G. Kinetic friction in or near the boundary region. II. The influence of sliding velocity and other variables on kinetic friction in or near the boundary region, 439.
Forrester, P. G. *See* Chalmers, Forrester & Phelps.
Free radical polymerization of the vapours of certain vinyl derivatives (Jones & Melville), 37.
Freudenthal, A. M. The statistical aspect of fatigue of materials, 416.

Gadsby, J., Hinshelwood, C. N. & Sykes, K. W. The kinetics of the reactions of the steam-carbon system, 129.
Grime, G. & Sheard, H. The experimental study of the blast from bombs and bare charges, 357.

Hinshelwood, C. N. *See* Gadsby, Hinshelwood & Sykes.

Jayaratnam Eliezer, C. On Dirac's theory of quantum electrodynamics: the interaction of an electron and a radiation field, 197.

Jayaratnam Eliezer, C. The application of quantum electrodynamics to multiple processes, 210.
Jones, T. T. & Melville, H. W. The free radical polymerization of the vapours of certain vinyl derivatives, 37.
Jones, T. T. & Melville, H. W. The photochemical polymerization of methyl vinyl ketone vapour, 19.

Kemball, C. The adsorption of vapours on mercury. II. The entropy and heat of adsorption of non-polar substances, 73.
Kemball, C. & Rideal, E. K. The adsorption of vapours on mercury. I. Non-polar substances, 53
Kinetic friction in or near the boundary region. I. Apparatus and experimental methods (Chalmers, Forrester & Phelps), 430.
Kinetic friction in or near the boundary region. II. The influence of sliding velocity and other variables on kinetic friction in or near the boundary region (Forrester), 439.
Kinetics of the reactions of the steam-carbon system (Gadsby, Hinshelwood & Sykes), 129.
Krishnan, R. S. The Raman spectrum of rock-salt, 188.
Krypton vacuum wave-length measurements (Littlefield), 220.

Lattice points in n-dimensional star bodies. I. Existence theorems (Mahler), 151.
Littlefield, T. A. Krypton vacuum wave-length measurements, 220.
Long, L. H. & Norrish, R. G. W. The thermochemistry of carbon: valence states, heats of sublimation and energies of linkage, 337.

McDowell, C. A. & Moelwyn-Hughes, E. A. Some thermodynamical properties of gaseous sulphur dichloride, 398.
Madhavarao, B. S., Thiruvenkatachar, V. R. & Venkatachaliengar, K. Algebra related to elementary particles of spin $\frac{3}{2}$, 385.
Mahler, K. On lattice points in n-dimensional star bodies. I. Existence theorems, 151.
Massey, H. S. W. *See* Bates and Massey.
Mechanism of ionospheric ionization (Woolley), 102.
Mechanism of ionospheric ionization, II (Woolley), 403.
Meksyn, D. Stability of viscous flow between rotating cylinders, I, 115.
Meksyn, D. Stability of viscous flow between rotating cylinders. II. Cylinders rotating in opposite directions, 480.
Meksyn, D. Stability of viscous flow between rotating cylinders. III. Integration of a sixth order linear equation, 492.
Melville, H. W. *See* Jones & Melville.
Moelwyn-Hughes, E. A. *See* McDowell and Moelwyn-Hughes.
Motion of waves in shallow water. Interaction between waves and sand bottoms (Bagnold), 1.

Norrish, R. G. W. *See* Long & Norrish.
Note on R. A. Bagnold's empirical formula for the critical water motion corresponding with the first disturbance of grains on a flat surface (Taylor), 16.

Phelps, E. F. *See* Chalmers, Forrester & Phelps.
Photochemical polymerization of methyl vinyl ketone vapour (Jones & Melville), 19.

Raman spectrum of rock-salt (Krishman), 188.
Rideal, E. K. *See* Kemball & Rideal.
Rochester, G. D. A cloud-chamber investigation of penetrating showers, 464.

Sheard, H. *See* Grime & Sheard.
Sneddon, I. N. The distribution of stress in the neighbourhood of a crack in an elastic solid, 229.
Stability of viscous flow between rotating cylinders, I (Meksyn), 115.
Stability of viscous flow between rotating cylinders. II. Cylinders rotating in opposite directions (Meksyn), 480.
Stability of viscous flow between rotating cylinders. III. Integration of a sixth order linear equation (Meksyn), 492.
Statistical aspect of fatigue of materials (Freudenthal), 416.
Sykes, K. W. *See* Gadsby, Hinshelwood & Sykes.

Taylor, Sir Geoffrey. A note on R. A. Bagnold's empirical formula for the critical water motion corresponding with the first disturbance of grains on a flat surface, 16.
Thermochemistry of carbon: valence states, heats of sublimation and energies of linkage (Long & Norrish), 337.
Thermodynamical properties of gaseous sulphur dichloride (McDowell & Moelwyn-Hughes), 398.
Thiruvenkatachar, V. R. *See* Madhavarao, Thiruvenkatachar & Venkatachaliengar.
Two-dimensional boundary-value problems of elasticity (Bibhutibhusan Sen), 87.

Venkatachaliengar, K. *See* Madhavarao, Thiruvenkatachar and Venkatachaliengar.

Woolley, R. v. d. R. The mechanism of ionospheric ionization, 102.
Woolley, R. v. d. R. The mechanism of ionospheric ionization, II, 403.

END OF THE ONE HUNDRED AND EIGHTY-SEVENTH VOLUME (SERIES A)

I.A.R.I. 75

INDIAN AGRICULTURAL RESEARCH
INSTITUTE LIBRARY, NEW DELHI.

Date of Issue	Date of Issue	Date of Issue
13.9.60		

GIPNLK—H-40 I.A.R.I.—29-4-55—15,000